INTRODUCTION

TO

ELECTRICITY AND OPTICS

INTRODUCTION

TO

ELECTRICITY AND OPTICS

BY

NATHANIEL H. FRANK, Sc.D.

Associate Professor of Physics

Massachusetts Institute of Technology

First Edition

Second Impression

McGRAW-HILL BOOK COMPANY, Inc.

NEW YORK AND LONDON

1940

THE MAPLE PRESS COMPANY, YORK, PA.

PREFACE

This book has been written primarily as a textbook for the use of those second-year students at the Massachusetts Institute of Technology who intend to pursue further studies in electrical engineering, physics, or both. These students have completed a year's course in calculus and one in mechanics and heat utilizing the author's text[1] and are simultaneously pursuing a second course in calculus. The goal to which this book aspires is a compact logical exposition of the fundamental laws of the electric and the magnetic field and the elementary applications of these laws to circuits, to a study of the electrical and magnetic properties of matter, and to the field of optics. The treatment is quantitative throughout and an attempt has been made to imbue the reader with a sound understanding of the fundamental laws (the Maxwell equations) and with the ability to apply them to many and varied phenomena without resorting to special formulas and methods. Thus ordinary circuit concepts and the ideas necessary for the understanding of optics are presented as natural consequences of the basic field equations.

In attempting to carry through the above program in an elementary text, it has been found expedient to depart widely from the usual elementary treatments. The book is essentially divided into two parts. The first half sets as its aim a systematic development of the fundamental laws of the electric and magnetic fields for empty space, confining the discussion of the electrical properties of matter to those of conductors. In this connection all four electromagnetic vectors, $\mathcal{E}$, D, B, and H, are introduced from the very outset; hence the Maxwell equations for empty space are presented in forms which are perfectly general, retaining their validity in the presence of material bodies. The second half encompasses the electric and magnetic properties of matter and is based essentially on the electron theory of matter. The study of electromagnetic waves in dielectrics leads smoothly to

[1] "Introduction to Mechanics and Heat," N. H. Frank, McGraw-Hill Book Company, Inc., New York.

the subject of optics; physical optics is largely emphasized, although one rather long chapter is devoted to geometrical optics.

Since it has become necessary to teach the m.k.s. system of units, these units have been employed from the very beginning along with the electrostatic system. Electromagnetic units are then introduced at an appropriate place. Thus the student learns the advantages of the newer m.k.s. system and at the same time becomes conversant with the older Gaussian units without which his understanding of much of the literature would be seriously handicapped. In view of the order of development of the subject matter, unrationalized units have been employed throughout and the transition to rationalized units, if desirable, may readily be made at a later stage of the student's education.

The first three chapters are devoted to the subject of the electrostatic field in vacuum. Here, as well as elsewhere in the book, the order of presentation of the topics has been chosen in accordance with the principle of introducing basic concepts one at a time wherever practicable. The concepts and definition of the magnetic field in empty space have been based solely on the mutual force actions of currents or of moving charges, care being taken to stress the magnetic induction vector B as the fundamental force vector. Discussion of magnetic poles is deferred to the second half of the book in connection with the properties of ferromagnetic media. The use of complex numbers is avoided in connection with simple a.c. circuits but the vector diagram method is derived from first principles. The two chapters which conclude the treatment for empty space introduce the concept of the Maxwell displacement current, a discussion of electromagnetic waves in free space, and the Poynting vector. Here the Maxwell equations are formulated in integral form to avoid the premature use of the symbolism of vector differentiation. Traveling and standing waves on an ideal transmission line serve to bridge the usual gap between oscillating LC circuits and the radiation field of an antenna, the latter topic being treated semi-quantitatively.

As previously mentioned, the second half of the book introduces the electrical and magnetic properties of matter on the basis of the Lorentz electron theory. Following a study of the essentially free electrons in vacuum tubes and in metals, the

properties of dielectrics in terms of bound electrons and of magnetic media in terms of the orbital motions and spins of the electrons are discussed. The difficult questions of mechanical forces on dielectrics and on magnetic bodies have been relegated to the problems in an attempt to develop critical methods of reasoning in this connection rather than a blind reliance on formulas. A dual standpoint has been adopted in connection with dielectrics and magnetic media: (1) the classical description in terms of dielectric constant and magnetic permeability and (2) the replacement of matter by equivalent charge and current distributions and the consequent reduction of a problem to one for empty space. It is gratifying to find that the treatment of magnetic problems in terms of Amperian currents is fully as simple as the equivalent method of introducing surface distributions of magnetic matter (magnetic poles) and has unquestionable pedagogical advantages. The laws of reflection and refraction of electromagnetic waves at dielectric boundaries are derived from the electromagnetic boundary conditions and the problem of intensity relations for normal incidence is discussed completely. An elementary, but quantitative, theory of the dispersion and scattering of light in gases is extended to give a physical picture of the nature of the refracted wave in isotropic and anisotropic media, the latter leading to the phenomena of double refraction. Fresnel diffraction is analyzed with the help of Fresnel zones and the Fraunhofer diffraction patterns of a single and of a double rectangular slit are worked out completely. The final chapter proceeds to a quantitative discussion of thermal radiation, including an elementary derivation of the law of radiation pressure and of the concept of electromagnetic momentum. A brief discussion of photometry and its connection with general radiation theory is also included.

A number of problems have been included at the end of each chapter. These problems have been designed not only to help the student learn how to apply fundamental principles to many and varied situations—and the working of many problems is essential for a thorough grasp of these fundamentals—but also in some cases to require the student to derive for himself a number of important general results not obtained in the text. A considerable range of complexity has been aimed at in these problems, and it is hoped that not too many of them can be solved

by use of formulas alone. It should be pointed out that this book is planned as a guiding textbook and, as such, should be supplemented by laboratory work and descriptive material, especially with regard to experimental methods.

The author would like to express his thanks to a number of his colleagues for valuable suggestions and criticisms and especially to John E. Meade for his able assistance in preparing the manuscript for publication.

N. H. FRANK.

CAMBRIDGE, MASS.,
May, 1940.

CONTENTS

INTRODUCTION TO ELECTRICITY AND OPTICS

CHAPTER I

THE ELECTROSTATIC FIELD OF FORCE

The subject of electrostatics deals with the laws governing the equilibrium interactions between "electrically charged" bodies or between "electrically charged" bodies and matter. It is assumed that the reader is acquainted with the qualitative facts concerning various methods of electrifying bodies, such as the historical method of rubbing a rod of hard rubber or bakelite with fur, as well as the more modern method of bringing the bodies into contact with one terminal of a battery or a power line. When bodies are so treated and placed near together, they exert forces on one another which did not exist before the treatment. We say that the bodies so treated have become electrically charged. For the present we shall not concern ourselves with the study of the methods of charging bodies electrically, but shall simply assume the possibility of so doing.

The fact that electrically charged bodies exert forces on each other when not in contact reminds one of gravitational forces and there is, in fact, a strong although not complete analogy between these two types of force. Just as in the case of gravitational forces, we can adopt one of two points of view with respect to these forces. First, there is the simple idea of "action-at-a-distance" forces and, secondly, the somewhat more involved concept of a field of force, produced, in this case, by the presence of electrified bodies. While it is true in studying electrostatics that both points of view are equally useful, the field concept proves to be ever so much more valuable in studying non-static phenomena and we shall adopt the idea of an electric field from the very outset. Suppose we have a single electrically charged body. We think of all space being affected by this body and

say that an electric field is established throughout space. If we place a second charged body at a given point of space, it will be acted on by a force (due of course in the last analysis to the presence of the first charged body) and we can use this force exerted on a test charge to describe the field of force. In this chapter we shall concern ourselves solely with a study of the nature of electrostatic fields of force and shall defer an examination of the sources of such fields until the next chapter.

1. Conservation of Electric Charge; the Electron. One of the fundamental facts concerning electrical forces is that, in contrast to gravitational forces, one may have both attractive and repulsive forces. If two bodies are brought into contact with the same terminal of a battery, it is found that they repel each other, whereas, if brought into contact with opposite terminals, they attract each other. This fact led to the assumption that there are two kinds of electricity, originally termed "vitreous" and "resinous" after the manner of their production, and now called "positive" and "negative," respectively. When a body is electrified, we say that *electric charge* has been transferred to the body, positive or negative charge as the case may be. There is a fundamental law concerning electric charge, namely, that no net electric charge can ever be created or destroyed. Whenever any positive charge is created, there is always created an equal amount of negative charge. The experimental proof of this law as given by Faraday will be discussed later. This law is known as the law of *conservation of charge* and is one of the most fundamental laws of physics.

It was postulated in the early days of the subject that electricity was a fluid, or rather two fluids, present in equal quantities inside matter, and that charging a body consisted of adding an excess of positive or negative fluid to it. There is overwhelming evidence today that electricity is atomic rather than continuous in nature and that the smallest electric charge available is that possessed by an *electron*, one of the fundamental particles of which atoms are composed. The electron carries a negative charge and its positive twin, the positron, has been recently discovered but does not possess the permanence of the negative electron. As one might expect, the charge carried by an electron is so exceedingly small compared to the ordinary charges with which one has to deal in large-scale experiments,

that for many purposes one may still think of electricity as being continuous rather than atomic.

To proceed to a quantitative formulation of the laws, we must introduce a measure of charge. There are two units of electric charge which will be of most use to us for the present, and these are, of course, arbitrarily defined. We postpone the actual definitions of these units to a later chapter but we shall name them and give their relative magnitudes. First, we have the *statcoulomb* or electrostatic unit of charge (e.s.u.). The other unit is the coulomb or practical unit which is a much larger unit. In fact, 1 coulomb $= 3.00 \times 10^9$ statcoulombs. The charge on an electron is 4.80×10^{-10} statcoulomb.

2. Intensity of the Electric Field.—The presence of an electric field in a given region of space can be detected by bringing into that region a so-called test charge, *i.e.*, a small *positively* charged body, and determining whether a force is exerted on this test charge. If such a force exists, we say that an electric field is present, and it would seem reasonable to specify the field by stating the force (direction and magnitude) exerted on the test body at each point of space. This procedure is open to a disturbing objection, namely, the force depends on the magnitude of the charge carried by the test body and hence cannot be used as a unique measure of the field. Hence we set up another measure of the strength of the field, the so-called *electric intensity*, which we denote by $\mathcal{E}$. The intensity is defined as a vector equal in magnitude to the *force per unit positive charge* exerted on the test body and of the same direction as this force. Thus, if the charge on the test body is q and at a given point of space there is an electrical force F exerted on it, the intensity $\mathcal{E}$ at that point of space is

$$\mathcal{E} = \frac{F}{q} \tag{1}$$

or

$$F = \mathcal{E}q. \tag{1a}$$

There still remains one question to be settled before we can be sure that $\mathcal{E}$ is a unique measure of the field strength. If the force exerted on the test body is proportional to the charge on the latter, then Eq. (1) provides a perfectly satisfactory definition. If, and this occurs in many cases, the force per unit

charge does depend on the size of the test charge due to the reaction of the latter on the sources of the field, we must generalize Eq. (1) to remove the ambiguity involved. Let us suppose that we bring a number of test charges successively to a given point of space and upon measurement we find that the ratio of force to charge does vary with the size of the test charge. We proceed to make the magnitude of the test charge smaller and smaller, the force becomes smaller and smaller, but the ratio of force to charge approaches a definite limiting value as the test charge is reduced indefinitely. This limiting value is then defined as the intensity of the field at that point. In symbols

$$\mathcal{E} = \lim \frac{\Delta F}{\Delta q} \text{ as } \Delta q \to 0 \qquad (1b)$$

Logically, we should adopt Eq. (1b) as the only strict definition of field intensity, but in many interesting cases the reaction of the test body is so small that Eq. (1) can be used as an excellent approximation.

A complete knowledge of a field involves the knowledge of the electric intensity, in both direction and magnitude at each point of space. Let us suppose that at each point of space there is constructed a vector representing the intensity of the field at that point. The totality of such vectors forms a vector field and is entirely similar to a gravitational field of force or to the vector field describing the motion of a fluid. Just as in these cases, we can construct lines of force or, more properly, lines of electric intensity which give the direction of the intensity at each point. Thus we can map out an electric field by use of these lines of force and then there remains only the task of specifying the magnitude of the intensity at each point. A special but important type of field is one in which the lines of force are all parallel straight lines and the intensity has the same magnitude at all points of the field. Such a field is called a *uniform* field and can exist only in a limited region of space. In general, however, the lines of $\mathcal{E}$ will be curved and $\mathcal{E}$ itself will vary in magnitude from point to point of space.

If we measure charge in statcoulombs (the electrostatic measure of charge) and force in dynes, intensity is then measured in *dynes per statcoulomb*. For reasons which will appear in the next section, this unit is sometimes called a *statvolt per*

centimeter. In the system of units involving the coulomb as unit charge, it turns out to be convenient to measure force in units of 10^5 dynes (this corresponds to using the kilogram as unit of mass, meter as unit of length, and second as unit of time) and corresponding intensity units are termed *volts per meter.*

3. Potential; Electromotive Force.—*The electrostatic field is a conservative field of force* just like the gravitational field. We postpone a formal proof of this statement until the next chapter but will examine some of the consequences of this fact in this section. If a test charge is moved around a closed path, the total work done on it by the electrostatic force is zero. Thus we see that the vector field can never be represented by lines of force which form closed loops. Furthermore, we can introduce the idea of *potential energy* of a test charge according to the usual definition that the gain of potential energy of the test charge in moving from one point in the field to another is the negative of the work done on it by the field during this motion. In other words, the work which we must do in pushing the charge can all be regained by allowing it to return to its initial position.

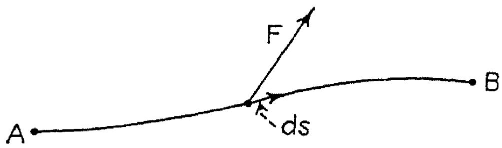

FIG. 1.

As in the preceding section, it is useful to refer all quantities such as work to unit charge, and we must introduce a few definitions. The work done by the field in moving a charge from point A to point B *per unit charge* is called the *electromotive force* along the path joining A and B. The work done in moving the charge q (Fig. 1) is, by definition,

$$W = \int_A^B F_s\, ds = q\int_A^B \mathcal{E}_s\, ds \tag{2}$$

using Eq. (1). F_s and $\mathcal{E}_s$ are the components of force and intensity in the direction of motion, respectively. Thus we have for the electromotive force along this path

$$\text{e.m.f.} = \frac{W}{q} = \qquad ds \tag{3}$$

Similarly, the potential energy per unit charge of a charge located at a given point is defined as the *electric potential,* or simply the *potential,* at that point. If we denote potential

energy by U and potential by V, we have for the gain of potential energy of the charge q as it moves from A to B (Fig. 1)

$$U_B - U_A = -W = -q\int_A^B \mathcal{E}_s\, ds \tag{4}$$

and the difference of electric potential of points A and B is

$$V_B - V_A = \frac{U_B}{q} - \frac{U_A}{q} = -\int_A^B \mathcal{E}_s\, ds \tag{5}$$

or the negative of the electromotive force along the path. If the charge gains potential energy in moving from A to B, we say that point B is at a higher electric potential than point A, and vice versa.

Furthermore, it is convenient to place the potential energy of our charge equal to zero at points infinitely distant from the region of space where the electric field exists (making the potential energy zero where the electrostatic force is zero) and thus have for the definition of the potential V at a point P,

$$V_P = -\int_\infty^P \mathcal{E}_s\, ds \tag{6}$$

In words, the potential at a point P is the work per unit charge which we must do on the charge to bring it from infinity to the point P *along any path whatsoever.* Equation (6) may be used only if the intensity at points distant from the sources drops off faster than $1/r$, where r is the distance from the sources to the field point. In any case, however, Eq. (5) is applicable.

The description of an electrostatic field in terms of the potential at each point instead of the intensity is decidedly simpler since the potential is a scalar quantity. We now have to show that one can deduce the intensity of a field at any point of space from a knowledge of the distribution of potential. Let us suppose that we know the potential V at each point of space, *i.e.*, that we are in possession of a relation of the form $V = f(x,y,z)$, where the function $f(x,y,z)$ depends on the particular type of field under discussion. If we wish to know what points of space are at the same potential V_0, we set $V = V_0$ and obtain an equation $f(x,y,z) = V_0 = \text{constant}$. This is the equation of a surface, and this surface is called an *equipotential* surface. There will exist a whole family of these equipotential surfaces corresponding to various values of V_0. Since, by definition, it requires no work

to move a charge from one point to another on the same equipotential surface, it follows that the lines of force (or intensity) must be at right angles to the equipotential surfaces at every point. Thus we have a connection between the lines of electric intensity and the potential. If one has a uniform field, for example, then the equipotentials consist of a set of parallel planes, the normals to these planes being the lines of force. Conversely, if one knows the equipotentials, one may immediately construct the lines of force by drawing the lines which intersect these surfaces at right angles. For example, suppose the equipotential surfaces consisted of a family of concentric spherical surfaces with a common center at a point O. Then the lines of force would consist of a set of straight lines radiating in all directions from O, since these lines, being radii of all the spherical equipotential surfaces, intersect the surfaces at right angles.

Suppose we consider a charge situated at a point P in an electric field and we move the charge a distance ds to a neighboring point. From Eq. (5) it follows that the change of potential dV is given by

$$dV = -\mathcal{E}_s\, ds \tag{7}$$

or

$$\mathcal{E}_s = -\frac{dV}{ds} \tag{8}$$

In words, the component of electric intensity in any given direction s is equal to the negative rate of change of potential with position along this direction. The space derivative in Eq. (8) is known as a *directional* derivative, since its value depends on the direction in which ds is taken. This checks the fact that $\mathcal{E}_s$, being the component of a vector, has a definite direction as well as a magnitude. If one moves from a point P to a neighboring point on the same equipotential, then dV/ds is zero in this direction since the potential does not change. If one moves, however, to a neighboring point not on the same equipotential, we obtain a value of dV/ds different from zero. The direction for which dV/ds has the *maximum* value possible at a given point is along a line of force, and the negative of this maximum rate of change of V with distance is the vector electric intensity at the point in question. This maximum rate of change of potential with position is known as the *gradient* of the potential, and this

is a vector pointing at right angles to the equipotential surface. In symbols we write

$$\mathcal{E} = -\operatorname{grad} V \tag{9}$$

An example may help clarify the above statements. Suppose one is walking on a hillside and is interested in the changes in elevation, vertical height from the bottom, as one walks. The points of equal elevation form curves called contours, and one often sees maps with these contour lines indicating the various elevations. These contour lines correspond to the equipotentials in the case of the electric field. Suppose one is standing at a point of 100-ft. elevation and wishes to descend to a point of 95-ft. elevation. There are many ways of doing this; all that is necessary is that one move from the starting point to any point on the 95-ft. contour. There is, however, a shortest way; this is along the perpendicular to the contours and is the direction of *maximum* slope, the line of steepest descent. It is clear that the maximum rate of change of elevation with position is along the line of steepest descent (or ascent) and this corresponds to the gradient of the elevation. In the electrostatic case, one moves from one equipotential surface to another differing in potential by a definite amount dV along the shortest path by following a line of force perpendicular to the surfaces. The rate of change of potential with distance in this direction is thus the gradient of the potential and is the maximum rate of change at the point in question.

The unit of potential or electromotive force in the electrostatic system of units is one erg per statcoulomb, usually called 1 *statvolt*. In the practical system of units, the unit of work is the joule = 10^7 ergs, so that the unit of potential or e.m.f. in this system is 1 joule/coulomb and is called 1 *volt*. One can easily show from the above that 1 statvolt = 300 volts.

4. Example.—In order to clarify further the principles presented in the preceding section, we shall apply them to a specific example. Suppose we are given the information that the distribution of electrostatic potential in the x-y plane is given by the equation

$$V \quad \frac{ax}{(x^2 +} \tag{10}$$

where a is a constant. (This is actually the potential of a tiny dipole at the origin, as we shall see in the next chapter.) We wish to determine the magnitude and direction of the electric intensity at any point in the plane and, if possible, derive the equation for the lines of electric intensity.

To obtain the x- and y-components of $\mathcal{E}$, we utilize the result expressed by Eq. (8). First, consider the change in potential as we move from a point, the coordinates of which are (x,y), to a neighboring point with coordinates $(x + dx,y)$. The displacement ds for this case is dx, and we thus obtain for the x-component of $\mathcal{E}$

$$\mathcal{E}_x = -\frac{\partial V}{\partial x} \tag{11}$$

In carrying out the indicated differentation, we must keep y constant, since x alone varies. Similarly, the y-component of $\mathcal{E}$ is given by

$$\mathcal{E}_y = -\frac{\partial V}{\partial y} \tag{12}$$

keeping x constant.

Using Eq. (10) for V, we find readily

$$\mathcal{E}_x = -\frac{a}{(x^2 + y^2)^{\frac{3}{2}}} + \frac{3ax^2}{(x^2 + y^2)^{\frac{5}{2}}} = \frac{a(2x^2 - y^2)}{(x^2 + y^2)^{\frac{5}{2}}} \tag{13}$$

and

$$\mathcal{E}_y = \frac{3axy}{(x^2 + y^2)^{\frac{5}{2}}} \tag{14}$$

The magnitude of $\mathcal{E}$ is then obtained from

$$\mathcal{E}^2 = \mathcal{E}_x^2 + \mathcal{E}_y^2$$

as

$$|\mathcal{E}| = \frac{a(4x^2 + y^2)^{\frac{1}{2}}}{(x^2 + y^2)^2} \tag{15}$$

which is not particularly simple but does allow a calculation of the field intensity at each point. There remains the question of the direction of the electric vector, *i.e.*, the direction of the lines of electric intensity. By definition the lines of $\mathcal{E}$ are drawn so that the tangents to these lines at every point give the direction of $\mathcal{E}$. Thus the tangent of the angle which the vector

$\vec{\mathcal{E}}$ makes with the x-axis is equal to the slope dy/dx of the lines of $\mathcal{E}$. Writing this as an equation, we have

$$\frac{dy}{dx} = \frac{\mathcal{E}_y}{\mathcal{E}_x} \tag{16}$$

and, using Eqs. (13) and (14), this becomes

$$\frac{dy}{dx} = \frac{3xy}{2x^2 - y^2} \tag{17}$$

If we can integrate this, the resulting equation is the equation for the lines of electric intensity.

Equation (17) is not easy to integrate as it stands, and indeed the whole problem looks rather formidable when handled in terms of Cartesian coordinates. For example, the equation for the equipotentials (more precisely, for the curves which represent the intersections of the equipotential surfaces with the x-y plane), which is obtained by setting $V = V_0$ (a constant) in Eq. (10), turns out to be of the sixth degree.

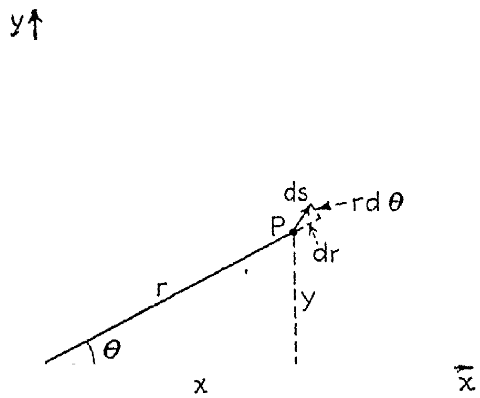

Fig. 2.

The whole problem becomes much simpler if we utilize polar coordinates in the plane, and we shall proceed to do so, checking the results already obtained and carrying the solution through to completion. In Fig. 2 the relations between the Cartesian coordinates x,y of point P and its polar coordinates r,θ are clearly brought out. We have

$$x = r \cos \theta; \qquad x^2 + y^2 = r^2$$

$$y = r \sin \theta; \qquad \frac{y}{x} = \tan \theta$$

Expressing Eq. (10) in terms of r and θ, we have

$$V = \frac{a \cos \theta}{r^2} \tag{18}$$

for the potential. The equipotentials are given by $r^2 = k \cos \theta$, ($k = a/V_0$), which is a much simpler equation to plot than Eq. (10). Now let us compute the radial and tangential compo-

nents of $\mathcal{E}$. For the radial component we have, according to Eq. (8),

$$\mathcal{E}_r = -\frac{\partial V}{\partial r} \tag{19}$$

since $ds = dr$, and for the tangential component

$$-\frac{1}{r}\frac{\partial V}{\partial \theta} \tag{20}$$

since for this case $= r\,d\theta$ (consult Fig. 2). Using Eq. (18), there follows

$$\mathcal{E}_r = \frac{2a \cos \theta}{r^3} \tag{21}$$

and

$$\mathcal{E}_\theta = a \sin \theta \tag{22}$$

The magnitude of $\mathcal{E}$ is obtained from $\mathcal{E}^2 = \mathcal{E}_r^2 + \mathcal{E}_\theta^2$, yielding

$$|\mathcal{E}| = \frac{a(1 + 3 \cos^2 \theta)^{\frac{1}{2}}}{r}$$

which is identical with Eq. (15). The proof is left to the student. The equations of the lines of $\mathcal{E}$ may be obtained as follows: The angle between $\mathcal{E}$ and the radius vector r is equal to the angle between an infinitesimal length ds of the curve and r (Fig, 3). Thus we may write

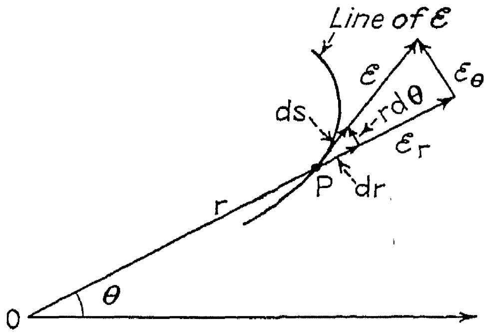

FIG. 3.

$$r\frac{d\theta}{dr} = \frac{\mathcal{E}_\theta}{\mathcal{E}_r} \tag{24}$$

and, using Eqs. (21) and (22), this becomes

$$\frac{d\theta}{dr} = \frac{1}{2}\frac{\tan \theta}{r} \tag{25}$$

If we rewrite this equation in the form

$$\frac{dr}{r} = 2\frac{d\theta}{\tan \theta} = 2\frac{\cos \theta\, d\theta}{\sin \theta} = 2\frac{d(\sin \theta)}{\sin \theta}$$

it may be integrated immediately, yielding

$$\ln\left(\frac{r}{\sin^2\theta}\right) = \text{constant}$$

or

$$r = b\sin^2\theta \tag{26}$$

where b is an arbitrary constant. The solid curves of Fig. 4 are plots of Eq. (26) for different values of b, and the dotted curves are a few of the equipotentials which, according to Eq. (18), are given by $\cos\theta/r^2 = \text{constant}$. The equipotentials intersect the lines of electric intensity at right angles. The fact that the lines of $\mathcal{E}$ are normal to the equipotentials allows the calculation of the latter from a knowledge of the former. Thus, for example, were we given Eqs. (21) and (22) as the initial data, we would write for the slope of the lines of force

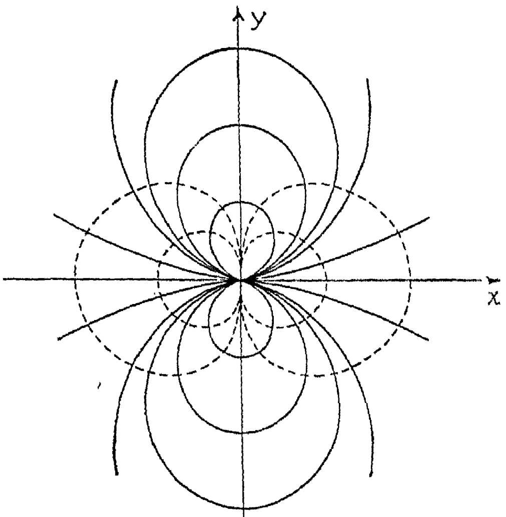

FIG. 4.

$$dr \qquad \frac{1}{2}\tan\theta$$

as in Eq. (25). At any point the slope of the curve intersecting a given curve at right angles is the negative of the reciprocal of the slope of the latter. Hence the equipotential curves satisfy the equation

$$1\,dr \qquad 1$$

or

$$\tan\theta\,d\theta + 2\frac{dr}{r} = 0$$

Upon integration this yields

$$\frac{\cos\theta}{r^2} = \text{constant}$$

which checks the potential as given by Eq. (18).

As a final check, let us rewrite Eq. (26) for the lines of $\mathcal{E}$ in Cartesian components, differentiate to find the slope and see if the result agrees with Eq. (17). Since $r = \sqrt{x^2 + y^2}$ and $\sin^2 \theta = y^2/r^2 = y^2/(x^2 + y^2)$, Eq. (26) becomes

$$\frac{(x^2 + y^2)^{\frac{3}{2}}}{y^2} = b \tag{26a}$$

and this is in the form

$$f(x,y) = \text{constant}$$

Differentiating, we have

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy = 0$$

so that

$$\frac{dy}{dx} = -\frac{\partial f/\partial x}{\partial f/\partial y}$$

From Eq. (26a) one finds

$$\frac{\partial f}{\partial x} = \frac{3x}{y^2}(x^2 + y^2)^{\frac{1}{2}}$$

and

$$\frac{\partial f}{\partial y} = \frac{3}{y}(x^2 + y^2)^{\frac{1}{2}} - \frac{2}{y^3}(x^2 + y^2)^{\frac{3}{2}} = \frac{y^2 - 2x^2}{y^3}(x^2 + y^2)^{\frac{1}{2}}$$

Hence

$$\frac{dy}{dx} = \frac{3xy}{2x^2 - y^2}$$

which is identical with Eq. (17).

To summarize, we have seen how one may find the components of the electric intensity if the potential is known and from this how to find an expression for the slope of the lines of force at any point. The latter expression, when integrated, leads to an equation (or equations) for the lines of electric intensity. Conversely, if the components of the field are known, the slope of the equipotentials (more precisely, the intersection of the equipotential surfaces with a plane) may be written down and, if integrated, yields the equation of the equipotentials. One may obtain the potential more directly, however, by utilizing

Eq. (6). In applying this equation, we have the freedom of choosing the path along which we integrate as we wish, and this may simplify the calculation enormously. For example, we may integrate along a radius vector r, starting at $r = \infty$ and ending the integration at the value of r corresponding to the point P at which we desire the potential. In this case, we have [using Eqs. (21) and (6)]

$$V = -\int_{\infty}^{r} \mathcal{E}_r \, dr = -2a \cos \theta \int_{\infty}^{r} \frac{dr}{r^3} = \frac{2a \cos \theta}{2r^2}$$

or

$$V = \frac{a \cos \theta}{r^2}$$

which checks Eq. (18).

In general, especially for three-dimensional problems, it is not feasible to attempt an integration of Eq. (16) or its equivalent in three dimensions and the lines of electric intensity are best obtained by graphical methods.

5. Metals as Equipotentials. We must now recall to the reader the fact that in a general way substances may be classified as conductors of electricity or as nonconductors or insulators. In our previous sections we have assumed tacitly that the bodies which were charged electrically were either insulators or, if conductors, were supported by insulators, so that the charge on them did not escape. It must be understood that the distinction between a conductor of electricity and an insulator is not absolutely sharp. From an atomic viewpoint, conductors of electricity are those forms of matter in which some or all of the electric charges of which the body is composed (electrons in the case of metals, ions in the case of aqueous solutions) can move more or less freely under the action of electric forces. In insulators, on the other hand, the electrons are held more or less rigidly fixed in the atoms of the substance.

In this section we shall concern ourselves with the behavior of good conductors such as metals in electrostatic fields. First, let us consider what happens when a metallic body (supported by an insulator) is charged. The charge transferred to the metal, being free to move, will flow through the metal and finally reach an equilibrium distribution. This equilibrium distribution has the following properties:

1. All the charge resides on the surface of the metal.
2. The charge distributes itself in such a manner that the surface of the conductor becomes an equipotential surface.
3. Every point inside the metal is at the same potential as the surface, so that the electric field is zero at every point inside the metal.

The fact that the charge resides on the surface is due to the mutual repulsion of the elements of charge so that they move as far away from each other as possible. Since the surface charge distribution is an equilibrium distribution, there can be no component of electric intensity parallel to the surface. Were this not so, the charge would be accelerated along the surface and equilibrium conditions would not prevail. Thus the electric intensity vector $\mathcal{E}$ just at the metal surface is normal to it, the lines of force starting from (or terminating on) the surface in a direction perpendicular to it. Hence the surface is an equipotential. Inside the metal we still have the normal content of electrons which are free to move. For equilibrium there must be no net force on these charges, and hence the field is zero everywhere inside the metal. In an insulator it is possible to have a potential gradient inside the body without flow of charge, but this cannot happen for a conductor.

The mobility of charge in a metallic conductor gives rise to complicated phenomena which one does not encounter in the case of gravitational fields. Suppose we bring an uncharged piece of metal into a region of space where an electric field exists. The mobile charges in the metal will be acted on by forces and will flow through the metal and distribute themselves in such a manner that the metal becomes an equipotential. Although the net charge on the metal is zero, there is a nonuniform distribution of charge set up on the surface, and the resultant of the field produced by these so-called *induced charges* and the original field is such as to make the whole metal an equipotential. Thus the presence of a metallic body in an electric field will, in general, modify the field considerably. This is true whether the metal be charged or uncharged.

One can make use of the above phenomena to charge metals by "induction" or "influence." Suppose a charged body is brought near an uncharged metallic body, as indicated in Fig. 5. There will be charges induced on the metal surface as indicated in the

figure, electrons being drawn closer to the positive external charge. If now the metal is connected to earth, the positive charge escapes, the negative charge being held by the attraction of the external positive charge. The connection to ground is now broken and then the external charge is removed. The metal is now found to be charged negatively, as one expects from the picture.

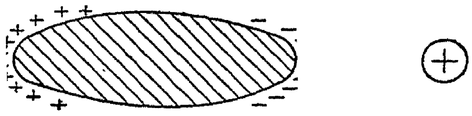

FIG. 5.

6. Motion of Charged Particles in Electric Fields.—The problem of determining the motion of a charged particle in an electrostatic field is essentially a dynamical one. Because the force acting is of electrical origin, the question of proper units is often troublesome, and we shall attempt to clarify this matter by solving a numerical example. We shall assume that the presence of the charged particle in the field does not appreciably disturb it, so that we may use Eq. (1) to calculate the force acting on the particle. If the particle carries a charge q and has a mass m, then the force acting on it is, according to Eq. (1),

$$F = \mathcal{E}q$$

and by Newton's second law

$$F = \mathcal{E}q = m\frac{dv}{dt} \tag{27}$$

where v is the vector velocity of the particle. If $\mathcal{E}$ is known as a function of position, the solution of the problem then follows the usual lines. Since electrostatic forces are conservative, we may apply the principle of conservation of mechanical energy. This takes the form

$$\frac{mv^2}{2} + qV = \text{constant} \tag{28}$$

since qV is the potential energy of the charge q when at a position where the potential is V. The constant is the constant total energy of the motion. Equation (28) is often useful in dealing with problems of the type under discussion.

If the charge q is expressed in statcoulombs (e.s.u.), then $\mathcal{E}$ must be in statvolts per centimeter (dynes per statcoulomb),

V in statvolts (ergs per statcoulomb), m in grams, time in seconds, and length in centimeters. This leads to a dynamical problem in c.g.s. units. It is recommended that the student use this system exclusively at the beginning and, if the charge is given in coulombs or potential in volts, that these quantities be reduced to the electrostatic system of units before making numerical substitutions.

Let us consider a very elementary problem. Suppose an electron of charge -4.80×10^{-10} statcoulomb and mass 9.0×10^{-28} gram starts from rest in a *uniform* field of intensity 100 volts/cm. How long does it take this electron to move 10 cm.; what are its velocity and kinetic energy at this time? We choose an origin at the initial position of the electron and an x-axis in the direction of the field (to the right, let us say). Since for a *uniform* field the intensity is everywhere the same, we have a constant force acting on the electron, and the problem is the simple one of motion with constant acceleration. From Eq. (27) this constant acceleration is

$$\frac{dv}{dt} = \frac{\mathcal{E}q}{m}$$

and, since the right-hand side is constant, this can be integrated immediately giving

$$\left.\begin{aligned} v &= v_0 + \frac{\mathcal{E}q}{m}t \\ \text{and} \qquad x &= v_0t + \frac{1}{2}\frac{\mathcal{E}q}{m}t^2 \end{aligned}\right\} \tag{29}$$

In our special case, $v_0 = 0$, since the particle starts from rest. The intensity $\mathcal{E}$ must be expressed in electrostatic units. Since 300 volts = 1 statvolt, we have

$$\mathcal{E} = \qquad \text{volts} \times \frac{1 \text{ statvolt}}{300 \text{ volts}} = \frac{1}{3} \text{ statvolt/cm.}$$

Using the second of Eqs. (29), we have, for $x = -10$ cm. (since the motion is opposite to the direction of the field due to the negative charge on the electron),

$$-10 = -\frac{1}{2} \cdot \frac{1}{3} \cdot \frac{4.8 \times 10^{-10}}{9.0 \times 10^{-28}} t^2$$

$$t^2 = \tfrac{9}{8} \times 10^{-16}$$

$$t = 1.06 \times 10^{-8} \text{ sec.}$$

From the first of Eqs. (29) we find for the velocity at this time

$$= -\frac{1}{3} \times \frac{4.8 \times}{9.0 \times 1} \quad \frac{0}{2} \times 1.06 \times 10^{-8} = -1.9 \times 10^{9} \text{ cm./sec.}$$

The kinetic energy at this point is

$$\text{K.E.} = \tfrac{1}{2}mv^2 = \tfrac{1}{2} \times 9.0 \times 10^{-28} \times (1.9)^2 \times 10^{18} = 16 \times 10^{-10} \text{ erg}$$

We can obtain the last result very quickly from energy considerations. Since the electron moves to the left, opposite to the direction of the field, it moves from a point of given potential to points of higher potential. Because of its negative charge, however, its potential energy decreases, and there must be a corresponding gain of kinetic energy. From Eq. (8) or Eq. (9) we have for the relation between intensity and potential for this case (field in the x-direction)

$$\mathcal{E} = -\frac{dV}{dx} \tag{30}$$

and, since $\mathcal{E}$ is constant, this yields

where V_1 is the potential at x_1 and V_2 the potential at x_2. For our case $x_1 = 0$ and $x_2 = -10$ cm., so that we have

$$V_2 - V_1 = \tfrac{1}{3} \times [0 - (-10)] = \tfrac{10}{3} \text{ statvolt} = 1{,}000 \text{ volts}$$

as the difference of potential between points x_1 and x_2. The change in potential energy of the electron in moving between these two points is, by definition,

$$U_2 - U_1 = q(V_2 - V_1) = -4.8 \times 10^{-10} \times \tfrac{10}{3} = -16 \times 10^{-10} \text{ erg}$$

and this is a decrease as expected. Thus the gain of kinetic energy is 16×10^{-10} erg, and, since the particle starts from rest, this is the kinetic energy after it has moved 10 cm. This checks the answer found above.

Problems

(The charge carried by an electron is $e = 4.80 \times 10^{-10}$ e.s.u., and its mass is 9.0×10^{-28} gram.)

1. The intensity of an electric field in the x-y plane is given by

$$\mathcal{E}_x = \frac{a(2x^2 - y^2)}{(x^2 + y^2)^{\frac{5}{2}}}$$

$$\mathcal{E}_y = \frac{3axy}{(x^2 + y^2)^{\frac{5}{2}}}$$

where a is a constant.

a. Compute the potential at a point P (coordinates x,y) by calculating the work per unit charge which must be done in moving a charge from infinity in along the x-axis to the point $(x,0)$ and then along a path parallel to the y-axis to the point P.

b. Repeat the calculation of part a along the following path; from infinity along the y-axis to the point $(0,y)$ and then parallel to the x-axis to the point P.

$$V = \frac{ax}{(x^2 + y^2)^{\frac{3}{2}}} \quad Ans.$$

2. The potential of a "dipole line" is given by

$$V = \frac{cx}{x^2 + y^2}$$

where c is a constant.

a. Show that the equipotential surfaces are circular cylinders.

b. Plot V as a function of x along the line $y = z = 0$.

c. Compute the components of $\mathcal{E}$ at any point, in both Cartesian and polar (cylindrical) coordinates.

d. From your answer to part c find the polar equation of the lines of electric intensity in the x-y plane.

e. Find expressions for the force which would be exerted on a charge q placed

(1) At the point $x = 1,\ y = z = 0$.

(2) At the point $x = z = 0,\ y = 1$.

3. The intensity of an electric field between two long concentric cylindrical surfaces is given by

$$\mathcal{E}_x = \frac{cx}{x^2 + y^2}, \qquad \mathcal{E}_y = \frac{cy}{x^2 + y^2}, \qquad \mathcal{E}_z = 0$$

where c is a constant and the x-y plane is perpendicular to the axis of the cylinders.

a. Construct the lines of force in the x-y plane, and sketch in some of the equipotentials.

b. Calculate the distribution of potential for this field, and show that the potential depends only on the distance from the axis.

c. Make a plot of potential against distance from the axis.

4. It is possible to produce the following potential in a limited region of space:

$$V = 2x^2 - y^2 - z^2$$

a. What are the components of the electric intensity in this field?

b. Find the equation of the lines of force in the plane $z = 0$, and plot the equipotentials and lines of force in this plane.

c. An electron is liberated at the point $x = z = 0$, $y = 1$. What sort of motion does it perform if it starts from rest?

d. Solve part *c* if the electron starts from rest at the point $x = 1$, $y = z = 0$.

5. An X-ray tube contains a filament where electrons are emitted with negligible velocities and a metallic target or anode placed some distance from the filament. The anode is maintained at a potential of 30,000 volts above the filament.

a. How fast are the electrons moving when they strike the anode?

b. How many electrons hit the anode per second if the electric current in the tube is 10 ma.? (1 amp. of current corresponds to 1 coulomb of charge reaching the anode per second.)

c. What is the force exerted on the anode due to this electron bombardment?

d. How much heat is generated per second?

6. A rigid insulating rod of length l, carrying charges $+q$ and $-q$, respectively, at each of its ends, is suspended by a thread attached to its center as a torsion pendulum in a uniform electric field of intensity $\mathcal{E}$, which is perpendicular to the suspending thread. Assume that the thread exerts no restoring torque when twisted.

a. Show that the resultant force on the rod is zero no matter what angle the rod makes with the field.

b. When in equilibrium what is the angle which the rod makes with the field?

c. If the rod is displaced from its equilibrium position by twisting it so that it makes an angle θ with the field, calculate the torque about the axis of suspension acting on the rod when in this position. What is the potential energy of the system in this configuration?

d. Considering the angle θ to be small, what sort of motion will the rod perform when released from the position described in part *c*? What will the period of the motion be?

7. The nucleus of a hydrogen atom may be considered a fixed mass point carrying a charge $+e$ equal to and opposite in sign to that of the electron. The potential due to this nucleus is $V = e/r$ (using e.s.u.) where r is the distance from the nucleus. An electron moves about this nucleus in a circle of radius a.

a. What is the force in dynes acting on the electron when in this orbit?

b. What must the speed of the electron be so that this motion is possible?

c. What is the total energy of the electron (kinetic plus potential)?

d. In the normal hydrogen atom it is found that the total energy of the electron in its orbit is -2.16×10^{-11} erg. What is the radius of the circle in which the electron moves?

8. A uniform electric field is set up between two metal plates *A* and *B* and a stream of electrons enters this field, as indicated in Fig. 6, with an initial velocity acquired by falling from rest through a potential difference of 1,000 volts. The plates are 2 cm. long and of $\frac{1}{2}$-cm. separation. The electrons fall on a fluorescent screen 30 cm. from the plates and give rise to a visible spot. If the difference of potential between the plates is 50 volts,

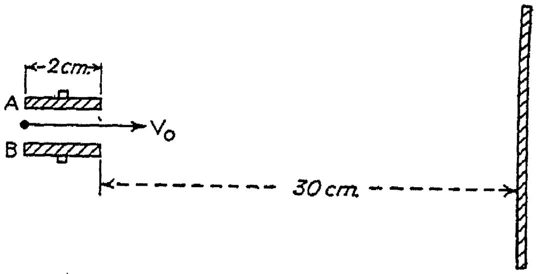

FIG. 6.

calculate the deflection of the spot from its position when no field exists between the plates.

9. A three-element vacuum tube may be idealized as follows: A plane metallic sheet (the filament) emits electrons with negligible velocities. Parallel to this filament is a plane grid of wires (the grid) separated by a distance of 2 mm. Another solid metallic sheet (the plate) is 10 mm. beyond the grid and parallel to it. Assume that uniform fields exist between filament and grid and between grid and plate.

a. What conditions must hold among the potentials of filament, grid, and plate so that electrons may reach the plate?

b. For given potentials of filament and plate how does the grid potential affect the velocity with which an electron strikes the plate?

c. Suppose the grid is made 10 volts positive and the plate 10 volts negative (both with respect to the filament). What will the motion of an electron be if it starts from the filament with zero velocity?

10. An oil drop of density 0.90 gram per cubic centimeter is held at rest by a vertical electric field of intensity 7.69 statvolts/cm. If the electric field is removed, it is found that the drop falls with a constant velocity of 1.09×10^{-2} cm./sec. Utilizing Stokes's law with the coefficient of viscosity of air equal to 1.8×10^{-4} c.g.s. unit, calculate

a. The radius of the drop.

b. The electric charge on the drop.

(This is the principle of the method utilized by Millikan to measure the charge on an electron.) For Stokes's law, consult Frank "Introduction to Mechanics and Heat."

CHAPTER II

THE SOURCES OF THE ELECTROSTATIC FIELD

We are now ready to undertake a more detailed discussion of the question of the production of electrostatic fields, inquiring into the laws which govern the dependence of such fields on the positions and magnitudes of the charges which produce them. At the very outset we encounter an experimental fact which, in sharp contrast to the corresponding facts for gravitational fields, introduces some complication. One finds experimentally that the intensity of an electric field produced, let us say, by a number of fixed charges depends, not only on the positions and magnitudes of these fixed charges, but also on the material medium in which these charges are embedded. For example, the force acting on a test charge in empty space is found to be reduced to a fraction of its original value when sources and test charge are immersed in a nonconducting liquid, such as oil. This dual dependence of intensity on both sources and medium makes it desirable from the standpoint of a logical, systematic procedure to develop a mode of description which shall, as far as possible, distinguish between these two effects.

In the course of our study we shall find that a complete separation of these two effects can be attained for the case of homogeneous media which are large enough so that the effects of the boundaries are negligible at all points at which we are interested in the field. (Under certain conditions the boundaries produce no effects even if in the immediate neighborhood of the field points.) We shall postpone a detailed examination of boundary effects (Chaps. X and XI) and for the present concern ourselves only with the case of *infinite homogeneous media*, and indeed principally with the case of empty space. In so doing, we shall be able to introduce a field vector different from $\mathcal{E}$ which shall depend *only* on the strengths and positions of the sources. We can then formulate independently the relation between the new vector and $\mathcal{E}$, this latter relation taking into account the dependence of the electric intensity on the medium in which the charges are immersed.

7. Coulomb's Law; the Electric Displacement Vector.—The law of force describing the interaction of electrical charges was formulated by Coulomb in 1785 as a result of experiments with a torsion balance. Coulomb found that the force of attraction or repulsion between two "point" charges was proportional to the products of the charges and inversely proportional to the square of the distance between them. The validity of Coulomb's law has been established with a high degree of precision, the result being obtained deductively from the fact that everywhere inside a hollow conductor the electric field is zero. The law is very much like Newton's law of gravitational attraction, and the student should constantly draw parallels between the laws which we shall examine and those governing gravitation.

In symbols we express the content of Coulomb's law by writing

$$F \sim \frac{q_1 q_2}{r^2} \tag{1}$$

where r is the separation of the two charges q_1 and q_2. From the field standpoint we can write for the intensity of the field *due to a single point charge* q

$$\mathcal{E} \sim \frac{q}{r^2} \tag{2}$$

where r now denotes the distance from q to the point where $\mathcal{E}$ is measured. In empty space or in most homogeneous insulating media $\mathcal{E}$ is directed along the line connecting q and the point where $\mathcal{E}$ is measured, away from or toward q, according to whether q is positive or negative. As we have indicated in the introductory paragraph, the proportionality factor in relation (2) depends on the medium, and we proceed now to define a new field vector, the so-called *electric displacement vector*, denoted by D, which will depend only on the magnitude and positions of the charges which give rise to the field.

For the case of a *single* point charge q the vector D at a point P located at distance r from q is defined by

$$D = \frac{q}{r^2} \tag{3}$$

and is directed along the line connecting q and the field point P either away from or toward q, according to whether q is positive or negative.

For the case of the field due to a number of point charges, we calculate the contribution to the resultant vector of each charge from Eq. (3) and then find the resultant D by *vector* addition. In symbols we may write

$$D = \sum \frac{q}{r^2} \qquad \text{(vector addition)} \tag{4}$$

If we have to calculate the field due to a continuous distribution of charge, we use an integral in Eq. (4) instead of the summation. Since we are adding vectors, however, we must proceed with caution. The method of calculation is the following:

1. Write an expression for the infinitesimal vector dD at a given field point due to an element of charge dq in accordance with Eq. (3).
2. Resolve this vector into components dD_x, dD_y, dD_z.
3. Calculate each component of D by integration, *e.g.*,

$$D_x = \int dD_x$$

4. Find the resultant D from its components.

We shall illustrate this procedure in detail in a later section.

The relation between the vectors D and $\mathcal{E}$ is usually written in the form

$$D = \epsilon\mathcal{E} \tag{5}$$

This equation is to be looked upon as a definition of ϵ, which is called the *inductive capacity* or *permittivity* of the medium in which we are calculating the field. For homogeneous isotropic substances ϵ is a constant, so that the vectors D and $\mathcal{E}$ have the same direction and differ only in magnitude. In the case of empty space, and this is the only case with which we shall concern ourselves for the present, we write Eq. (5) as

$$D = \epsilon_0\mathcal{E} \tag{6}$$

where ϵ_0 denotes the *inductive capacity* or *permittivity* of empty space.

For the special case of a *single* point charge in empty space, we combine Eqs. (6) and (3) to obtain

$$\mathcal{E} = \frac{1}{\epsilon_0}\frac{q}{r^2} \tag{7}$$

and this is one analytic formulation of Coulomb's law for vacuum.

Before proceeding further, we must stop to consider the question of units. We have complete freedom in our choice of both the units and dimensions of ϵ_0 (or ϵ), and different choices lead to different systems of units. As previously indicated, we shall confine our attention in this discussion to two sets of units. In the *electrostatic system* of *units* (e.s.u.), we place ϵ_0 equal to unity and *without* dimensions (a pure number). Although this procedure has the distinction of being the first to be employed, it is not without its disadvantages. All distinction between D and $\mathcal{E}$ in vacuum is lost, and the vectors become identical. This identity does not exist in material media although ϵ is dimensionless. If we write Coulomb's law for the force between two point charges in vacuum, we have

$$F = \frac{q_1q_2}{\epsilon_0 r^2} \tag{8}$$

and if ϵ_0 is the pure number unity, the dimensions of charge can be expressed in terms of those of mass, length, and time. This is, as we have shown, quite an arbitrary procedure. *The unit charge in the electrostatic system, the statcoulomb, is defined so that two such similar charges separated by a distance of* 1 *centimeter in vacuum repel each other with a force of* 1 *dyne.*

The other system of units which we shall employ, the so-called m.k.s. system in which mass is measured in kilograms, length in meters, and time in seconds, differs from the e.s.u. in two ways: (1) the sizes of the units are different and (2) no attempt is made to express all quantities in terms of mass, length, and time. In fact, we treat electric charge as a new fundamental physical quantity having its own dimension, just as mass is introduced in mechanics. Thus the units of all quantities are to be expressed in terms of units of *mass, length, time,* and *charge.* As unit of charge in the m.k.s. system we have the *coulomb*, and at present we must content ourselves with the fact that this is equal to 3×10^9 statcoulombs. The actual definition of the coulomb is not based on electrostatic laws, so that we must defer it to a

later chapter. We are, however, in a position to calculate the numerical value of ϵ_0 in the m.k.s. system. From Eq. (8) we see readily that ϵ_0 has the dimensions of q^2t^2/ml^3. Suppose we have two point charges, each 1 statcoulomb, placed 1 cm. apart in vacuum. The force is 1 dyne, and we can use Eq. (8) to obtain ϵ_0. Substituting numerical values in Eq. (8), there follows

$$10^{-5} = \frac{\frac{1}{9} \times 10^{-18}}{\epsilon_0 \times 10^{-4}}$$

or

$$\epsilon_0 = \tfrac{1}{9} \times 10^{-9}\ (\text{coulombs})^2(\text{sec.})^2/\text{kg.-meter}^3$$

In making this substitution we have used the facts that the unit force in the m.k.s. system is 10^5 dynes, that 1 cm. $= 10^{-2}$ meter and that 1 statcoulomb $= \frac{1}{3} \times 10^{-9}$ coulomb.

8. Gauss's Law; Flux of D.—Just as in the case of the intensity vector $\mathcal{E}$, one can map the field of the vector D, constructing lines of D which give its direction at every point. In vacuum (and also in isotropic homogeneous media) the directions of D and $\mathcal{E}$ coincide at every point so that a single diagram represents both fields. As we have pointed out, the lines of intensity or displacement merely give information as to the directions of the vectors but give no information as to their magnitudes. We can insert this latter information by limiting the number of lines which we draw in a definite manner. The convention to be used is that the number of lines per unit area which pass through an element of area normal to these lines shall be made equal to the number of units in the magnitude of the vector at the point where the element of area is located. Thus, for example, if we consider a point where the intensity of the electric field is 10 units, we should (using the electrostatic system of units) construct 10 lines crossing 1 cm.^{2} which is normal to these lines at the field point.

Suppose we construct the lines of D due to a single point charge q in empty space according to the above scheme. These lines are straight lines radiating in all directions from q. At a distance r from q we construct a spherical surface of radius r and center at q (this is an equipotential), and this surface is everywhere normal to these lines. According to Eq. (3), D has the same value everywhere on this surface, and hence the number of lines per unit area crossing this surface is the same at every

point. Since there are q/r^2 lines per unit area and the area of the sphere is $4\pi r^2$, we see that the total number of lines of D crossing the surface is $4\pi q$. The total number of lines crossing any surface is called the *flux* across that surface, and we denote it by the symbol ψ. Thus we have $\psi = 4\pi q$. This flux does not depend on the radius of the spherical surface enclosing q and hence is the same for all such spherical surfaces, no matter what their radii. Thus we see that lines of D cannot start or stop in empty space but must terminate on charges. It is clear from the above that, if any closed surface is constructed which completely surrounds q, the total flux of D across it must

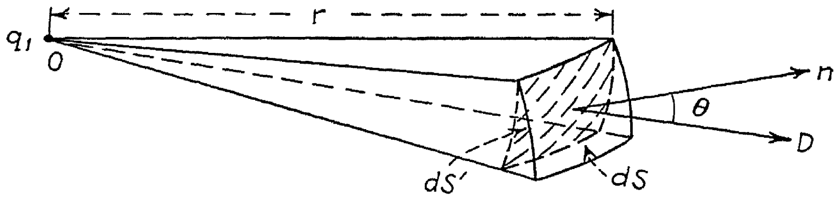

FIG. 7.

be $4\pi q$. When expressed formally as an equation, this is the content of Gauss's theorem. It follows immediately that the flux of $\mathcal{E}$ is $1/\epsilon_0$ times as great as the flux of D in empty space.

The formal statement of Gauss's theorem is as follows. *The surface integral of the normal component of the electric displacement taken over any closed surface is equal to $4\pi q$, where q is the total charge enclosed by the surface.* In symbols

$$\psi = \int D_n \, dS = 4\pi q \tag{9}$$

where D_n is the component of D along the normal n to the surface. We shall take the direction of the normal as positive if it points *outward* from the closed surface. To obtain a formal proof of Eq. (9), we imagine a charge q_1 at a given point O surrounded by an arbitrary closed surface. Now construct an infinitesimal cone with its vertex at O, intersecting the enclosing surface in an element of area dS at a distance r from O (Fig. 7). Let dS' be the projection of dS on a sphere of radius r. The flux of D across dS is

$$\begin{aligned} d\psi &= D_n \, dS = D \cos\theta \, dS \\ &= q_1 \frac{dS \cos\theta}{r^2} \end{aligned}$$

using Eq. (3).

Since $dS' = dS \cos\theta$ and dS'/r^2 is the solid angle $d\Omega$ subtended at O by dS, we have

$$d\psi = q_1\, d\Omega$$

and, if we integrate over the whole closed surface, we obtain

$$\psi = q_1 \int d\Omega = 4\pi q_1$$

since the solid angle subtended by any closed surface at a point inside it is 4π.

If the surface encloses a number of charges $q_1, q_2, q_3, \cdots, q_i, \cdots$, each charge gives rise to a flux $4\pi q_i$, so that the total flux across the surface becomes $\sum_i 4\pi q_i = 4\pi q$, where q is the total enclosed charge.

Although we have derived Gauss's law from the definition of D as given by Eq. (3) or by Eq. (4), it is clear from the principle of conservation of charge that it is perfectly general. On the other hand, Eqs. (3) and (4) are subject to the restrictions discussed in the introductory paragraph. *Gauss's law is one of the basic laws of the theory of electricity.*

9. Applications of Gauss's Law.—Gauss's law provides a simple means of calculating the field due to certain symmetrical charge distributions. The power of Gauss's law lies in the fact that we are free to apply it to any surface whatsoever. In those cases where the directions of the lines of D are known from considerations of symmetry, we can construct a surface so shaped that the evaluation of the integral $\int D_n\, dS$ becomes much simplified.

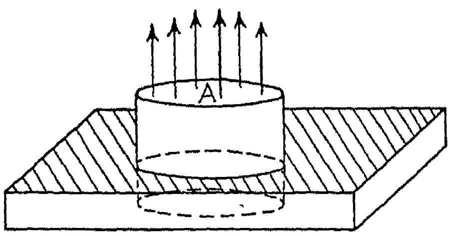

FIG. 8.

The following examples will illustrate the procedure.

a. Field of an Infinite Plane Metal Plate, Uniformly Charged.—In this case the symmetry requires that the lines of D be straight lines normal to the plate and the magnitude of D can depend only on the distance of the field point from the charged surface. Hence we construct an appropriate Gaussian surface in the form of a cylinder whose axis is normal to the plate, the top face lying above the surface and the bottom face lying inside the metal (Fig. 8). The curved surface has no flux passing through it,

since D is everywhere parallel to the cylinder axis. At the bottom face, D is everywhere zero, since no electric field can exist inside the metal. The total flux emerging from the cylinder comes through the top face where D is normal to the surface and has the same value at all its points. Thus we have

$$\psi = \int D_n \, dS = D \int_{\substack{\text{top}\\\text{face}}} dS = DA$$

where A is the area of the top face. By Gauss's theorem this must equal $4\pi q$, where q is the total charge inside the cylinder. Denoting the *charge per unit area by* σ (the surface-charge density), we have $q = \sigma A$ and hence

$$\psi = DA = 4\pi\sigma A$$

$$D = 4\pi\sigma \tag{10}$$

and is constant, independent of the position of the area A. The corresponding electric intensity is also constant in free space and has the value

$$\mathcal{E} = \frac{4\pi\sigma}{\epsilon_0} \tag{11}$$

This is an example of a uniform field. In practice one cannot have an infinite metal sheet, but for a plate of finite size the field is practically uniform at points whose distances from the plate are small compared to its surface dimensions.

The field intensity just at the surface of any charged conductor, no matter what its shape or size, is given by Eq. (11), where σ, the surface-charge density, will vary from point to point on the conducting surface. The lines of D and $\mathcal{E}$ leave the surface at right angles to it and start out at each point with the magnitude given by Eqs. (10) and (11), respectively. The proof of this statement is identical with the one we have given, with the single modification that one considers a cylinder of infinitesimal cross section and altitude, rather than one of finite dimensions. This is necessary since the charge is not uniformly distributed over the surface and the field is normal to the surface only in the immediate neighborhood thereof.

b. Field of a Long Uniformly Charged Straight Wire.—In this case the symmetry is such that the lines of D are all normal

to the axis of the wire. At a given field point P the value of D cannot depend on the x-coordinate of P (along the wire), since the wire is very long, nor on the angular position of P around the wire because of cylindrical symmetry. Thus the magnitude of D can only depend on the distance r of the field point from the axis of the wire. We therefore construct a closed surface in the form of a cylindrical can (Fig. 9) with radius r and altitude h. No flux crosses the top or bottom face of the cylinder, as the normal component of D is zero everywhere on these faces. On the curved surface, D has everywhere the same magnitude (since it depends only on r) and is normal to the surface at all its points. Thus the flux emerging from the cylinder is

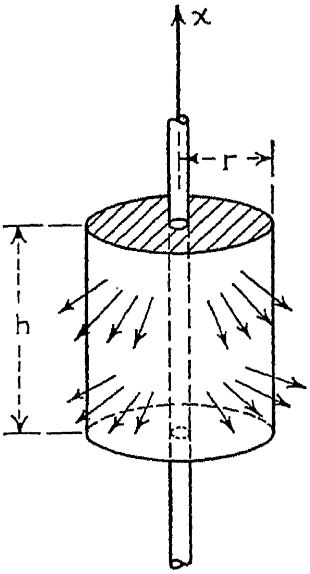

FIG. 9.

$$= \int_{\substack{\text{curved}\\ \text{surface}}} D_n \, da = 2\pi r h D$$

where D is the value of the displacement vector at a distance r from the axis of the wire. If the *charge per unit length* of the wire is τ, the total charge inside the cylinder is τh and Gauss's theorem requires that

$$\psi = 2\pi r h D = 4\pi h \tau$$

$$D = \frac{2\tau}{r} \qquad (12)$$

The corresponding electric intensity in empty space becomes

$$\mathcal{E} = \frac{2\tau}{\epsilon_0 r} \qquad (13)$$

and we see that the field strength drops off inversely as the distance from the wire. Contrast this with the field of a *single* point charge.

10. The Field of Fixed Charge Distributions. When one is faced with the problem of determining the field of an arbitrary distribution of charge, the direct application of Gauss's theorem as given in the preceding section is not convenient, and one can proceed by the method outlined in Sec. 6. For the case of the

field due to a number of point charges, the method requires a straightforward vector addition of the fields due to the individual point charges [utilizing Eq. (4)] and requires no further comment. As we have stated, the field due to a continuous distribution of charge can be calculated by an integration, and we shall illustrate the procedure with the help of examples.

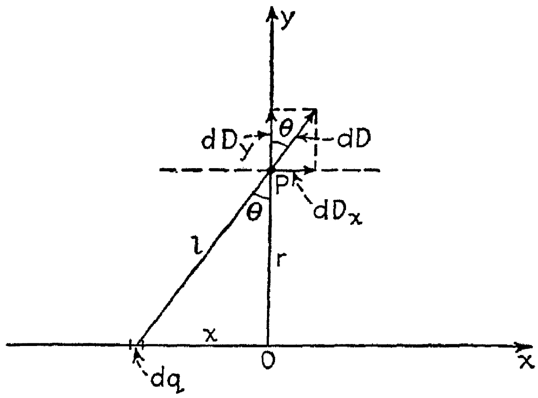

FIG. 10.

Let us calculate the field due to a long, uniformly charged straight wire by this method. Let the plane determined by the wire and the point P at which we are calculating D be the x-y plane, and construct x-y axes, as shown in Fig. 10. The contribution to D due the element of charge dq is shown as dD in the figure. If the charge per unit length of the wire is τ, we have

$$dq = \tau\, dx$$

The magnitude of dD is given according to Eq. (3) as

$$dD = \frac{\tau\, dx}{l^2} \tag{14}$$

We must next calculate the components dD_x and dD_y of this vector before integrating. These are

$$dD_x = dD \cdot \sin\theta = \tau \frac{dx \sin\theta}{l^2}$$

$$dD_y = dD \cdot \cos\theta = \tau \frac{dx \cos\theta}{l^2}$$

The components D_x and D_y of the resultant field at P are then

$$\left.\begin{aligned} D_x &= \tau \int_{-\infty}^{+\infty} \frac{dx \sin\theta}{l^2} \\ D_y &= \tau \int_{-\infty}^{+\infty} \frac{dx \cos\theta}{l^2} \end{aligned}\right\} \tag{15}$$

From symmetry we see without calculation that the first integral will vanish, so that the vector D is normal to the wire. The

integrations are most easily carried out using θ as the independent variable. We have (Fig. 10)

$$x = r\tan\theta; \qquad dx = r\sec^2\theta\, d\theta$$
$$r = l\cos\theta; \qquad \frac{1}{l^2} = \frac{\cos^2\theta}{r^2}$$

so that

$$\frac{dx}{l^2} = \frac{d\theta}{r}$$

and the integrals become

$$D_x = \frac{\tau}{r}\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} \sin\theta\, d\theta = \frac{\tau}{r}\Big| -\cos\theta \Big|_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} = 0$$

and

$$D_y = \frac{\tau}{r}\int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} \cos\theta\, d\theta = \frac{\tau}{r}\Big| \sin\theta \Big|_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} = \frac{2\tau}{r}$$

Thus we have $D = 2\tau/r$ and $\mathcal{E} = 2\tau/\epsilon_0 r$, coinciding with the results found utilizing Gauss's theorem and given by Eqs. (12) and (13).

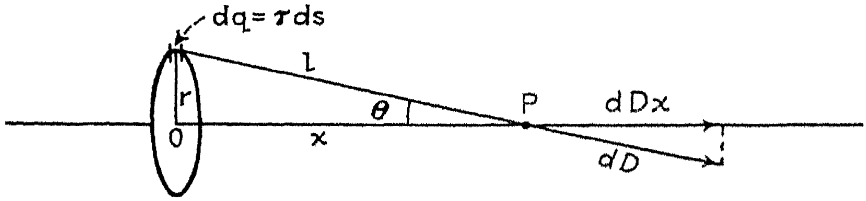

FIG. 11.

As a second example we shall calculate the field due to a uniformly charged circular ring of wire of negligible cross section at a point on the axis of the ring. Let the x-axis be the axis of the ring and the origin at the center of the ring, as shown in Fig. 11. Consider an element of charge dq on an element ds of the ring. This is

$$dq = \tau\, ds = \frac{q}{2\pi r} ds$$

where q is the total charge on the ring and r its radius.

At the point P the vector dD due to dq has the direction shown, and has the magnitude

$$dD = \frac{dq}{l^2} = \frac{q}{2\pi r} \cdot \frac{ds}{l^2} \tag{16}$$

If we consider all the vectors due to the various elements of charge dq on the ring, we see that they form a conical array of vectors with the apex at P and furthermore that they all have the same magnitude according to Eq. (16). If we resolve these vectors into components, it is evident that only the component dD_x yields anything to the resultant field, the other components adding up to zero. Thus we have

$$dD_x = dD \cos \theta = \frac{q}{2\pi r} \cdot \frac{ds}{l^2} \cdot \frac{x}{l}$$

and hence

$$D_x = \int\limits_{\substack{\text{around}\\ \text{ring}}} \frac{qx}{2\pi r} \cdot \frac{ds}{l^3} = \frac{qx}{2\pi r l^3} \int ds = \frac{qx}{l^3}$$

since $\int ds$ is simply the circumference of the circle. Thus we have the resultant field directed along the x-axis at the point P and equal to

$$D = \frac{\cdots}{l^3} = q\frac{\cdots}{(r^2 + \cdots} \tag{17}$$

with the corresponding value of $\mathcal{E}$ as

$$\mathcal{E} = \frac{qx}{l^3} = \cdots\frac{x}{\cdots + \cdots} \tag{17a}$$

Equations (17) and (17a) are valid only for points on the axis Ox, as is clear from the derivation.

11. The Use of Potential in Field Calculations.—The method of the preceding section, although straightforward, is cumbersome even in relatively simple problems because of the necessity of dealing with vector summations or integrations. The introduction of the electric potential, a scalar quantity, turns out to simplify the problem of the calculation of electrostatic fields considerably. In this section we shall investigate the use of the electric potential in this connection.

Our first task is to prove that the electrostatic field is conservative and hence that a potential exists. We start by considering the field of a point charge in empty space. Suppose a test charge is moved from one point A to another B in this field. We have to show that the work done on this charge, or that the electromotive force along the path, depends only on the starting and end points A and B and not on the path. Consider a portion of the path ds as shown in Fig. 12. The electromotive force along this elementary path is, by definition,

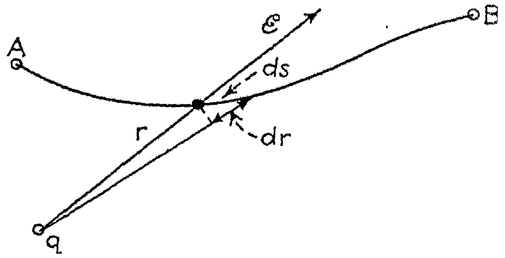

FIG. 12.

$$d\ (\text{e.m.f.}) = \mathcal{E}_s\, ds = \mathcal{E}\, ds \cos \angle_{ds}^{\mathcal{E}}$$

From the figure it is clear that

$$ds \cos \angle_{ds}^{\mathcal{E}} = dr$$

so that we have

$$d\ (\text{e.m.f.}) = \mathcal{E}\, dr = \frac{q}{\epsilon_0 r^2}\, dr$$

utilizing Eq. (7) for $\mathcal{E}$. For the whole path from A to B, we have

$$\text{e.m.f.} = \int_{r_A}^{r_B} \frac{q}{\epsilon_0 r^2}\, dr = -\frac{q}{\epsilon_0}\left(\frac{1}{r_B} - \frac{1}{r_A}\right) \tag{18}$$

as the electromotive force (the work per unit charge) along the path from A to B. Since this depends only on the positions of A and B (more precisely, on r_A and r_B), it is independent of the path connecting them. This shows that the field of a *single* point charge is conservative. It immediately follows that the field of an arbitrary distribution of charge is also conservative, since such a field can be obtained by superposing the fields of point charges and the e.m.f. can be calculated as the sum of the e.m.f.s. due to each point charge. Since each term of the sum is inde-

pendent of the particular path followed, then the whole sum is also independent of the path. This completes the proof.

We now investigate the potential of various charge distributions.

a. Potential of a Point Charge.—From Eq. (18) and the definition of potential difference as given by Eq. (5) of Chap. I, it follows that the difference of potential between two points A and B in the field of a *single* point charge q is given by

$$V_B - V_A = \frac{q}{\epsilon_0}\left(\frac{1}{r_B} - \frac{1}{r_A}\right) \tag{19}$$

Using the convention that the potential at points infinitely far from q be taken as zero [compare with Eq. (6) of Chap. I], we find for the potential at any point P of the field

$$V_P = \frac{q}{\epsilon_0 r} \tag{20}$$

where r is the distance from q to the point P. As stated before, this is the work per unit charge we must do on a test charge to move it from infinity to the point P against the repulsion of the charge q. The points lying on a sphere of radius r all are at the same potential and thus form an equipotential surface. The lines of D or $\mathcal{E}$ are normal to these surfaces and hence are the radii of such spheres. The gradient of the potential at any point is directed radially toward or away from q, and we have

$$\mathcal{E} = -\operatorname{grad} V = -\frac{dV}{dr} = \frac{q}{\epsilon_0 r^2} \tag{21}$$

which checks Eq. (7).

One can demonstrate readily, with the help of Gauss's theorem, that for the field of any *spherically symmetrical* distribution of charge, *e.g.*, a uniformly charged metal sphere, Eqs. (20) and (21) hold for all points *outside* the charge. For a proof of this involving a direct integration, see Frank "Introduction to Mechanics and Heat," Chap. XII. Thus a spherical charge distribution creates a field external to itself which is the same as if the charge were concentrated at the center of the sphere.

b. The Potential and Field of a Dipole.—By a *dipole* is meant a pair of equal and opposite point charges separated by a definite distance. We shall investigate the field of a dipole, charges $\pm q$, separation $2a$. Since there is symmetry about the axis of the dipole (the line connecting the charges), it will be sufficient to

restrict our attention to the plane containing the field point P and the dipole (Fig. 13).

The potential at P due to the charge $+q$ is $+\dfrac{q}{\epsilon_0 r_1}$ and that due to the charge $-q$ is $-\dfrac{q}{\epsilon_0 r_2}$. *Since potential is a scalar quantity, the potential at P is the algebraic sum of the two terms.* Thus we have at the point P

$$V = \frac{q}{\epsilon_0 r_1} - \frac{q}{\epsilon_0 r_2} = \frac{q}{\epsilon_0}\left(\frac{1}{r_1} - \frac{1}{r_2}\right) \tag{22}$$

In order to express V in terms of x and y, the coordinates of P, we use the relations

$$r_1^2 = (x - a)^2 + y^2; \qquad r_2^2 = (x + a)^2 + y^2$$

and Eq. (22) becomes

$$V = \frac{q}{\epsilon_0}\left[\frac{1}{\sqrt{(x - a)^2 + y^2}} - \frac{1}{\sqrt{(x + a)^2 + y^2}}\right]$$

The components of $\mathcal{E}$ may then be obtained by differentiating Eq. (23) with respect to x and y, respectively. In symbols we have, in accordance with Eq. (8) of Chap. I,

$$\partial x; \qquad \partial y$$

This example illustrates the general procedure: (1) Calculate the potential as an algebraic sum of the potentials of point charges. For continuous distributions this becomes an ordinary integration. This is always an easier process than the corresponding integrations in the method of the previous section because potential is a scalar quantity and there is no question of taking components.

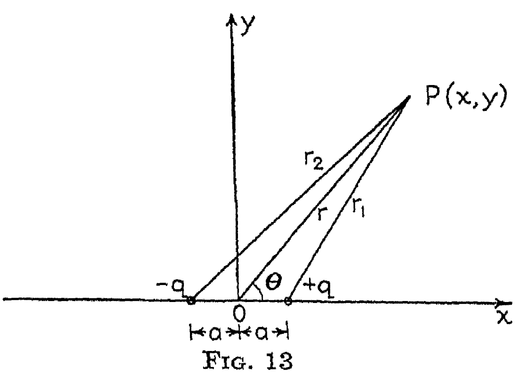

FIG. 13

(2) When the potential has been found, the intensity $\mathcal{E}$ can be computed by differentiation in accordance with the relation $\mathcal{E} = -\operatorname{grad} V$.

In the special but very important case where the distance $2a$ between the charges is small compared to r_1 and r_2, we can simplify Eq. (22) as follows. We write:

$$V = \frac{q}{\epsilon_0}\left(\frac{r_2 - r_1}{r_1 r_2}\right) \cong \frac{q}{\epsilon_0 r^2}(r_2 - r_1)$$

where r is the distance from the dipole to the field point. $(r_2 - r_1)$ is the difference in distance between the ends of the dipole and the point P and, if the charges are close together, we can write very nearly (Fig. 14)

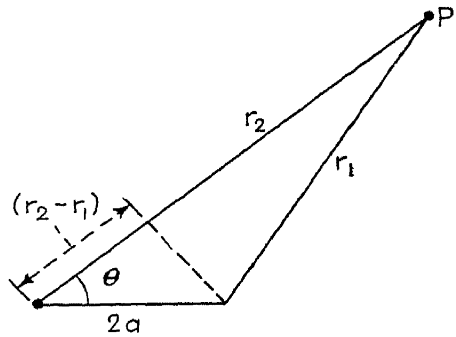

FIG. 14.

$$r_2 - r_1 = 2a \cos \theta$$

so that we obtain for V

$$V = 2aq \frac{\cos \theta}{\epsilon_0 r^2} \tag{24}$$

The expression $2aq$ is the product of the separation of the charges and the magnitude of either of them. This is called the *dipole moment* of the dipole, and we denote it by p. Using the symbol p, Eq. (24) becomes

$$V = \frac{p \cos \theta}{\epsilon_0 r^2} \tag{25}$$

as the potential of a tiny dipole of moment p. In terms of the x- and y-coordinates of the field point P the equation becomes, since $r^2 = x^2 + y^2$ and $\cos \theta = x/r$,

$$V = \frac{px}{\epsilon_0 r^3} = \frac{px}{\epsilon_0 (x^2 + y^2)^{\frac{3}{2}}} \tag{26}$$

from which $\mathcal{E}_x$ and $\mathcal{E}_y$ may be conveniently found. Further discussion of this field is left to the problems.

c. Potential of a Circular Ring of Charge.—To illustrate the case of fields due to a continuous distribution of charge, we consider the example of the previous section and calculate the potential of a uniformly charged ring at a point on the axis of the ring. In general the potential is given by

$$V = \frac{1}{\epsilon_0}\int \frac{dq}{r} \tag{27}$$

where r is the distance between dq and the field point. For dq we write $\tau\, ds$, $\sigma\, dS$, or $\rho\, dv$, depending on the problem at hand, where τ, σ, ρ are linear, surface, and volume charge densities, respectively.

Referring to Fig. 11, we have, in place of Eq. (27),

$$V_P = \frac{1}{\epsilon_0}\int \frac{dq}{l} = \frac{1}{\epsilon_0}\int \frac{\tau\, ds}{l} = \frac{1}{\epsilon_0 l}\int \tau\, ds = \frac{q}{\epsilon_0 l}$$

since l is the same for all the elements of charge dq on the ring. This can be written as

$$V = \frac{q}{\epsilon_0\sqrt{x^2 + r^2}} \tag{28}$$

Equation (28) gives the variation of the potential with the coordinates of points on the x-axis only and should not be applied at other points. We can, however, find the x-component of $\mathcal{E}$ from Eq. (28) for a point such as P. We have

$$\mathcal{E}_x = -\frac{dV}{dx} = \frac{qx}{\epsilon_0(r^2 + x^2)^{\frac{3}{2}}}$$

which is identical with Eq. (17a).

Problems

1. Two equal and opposite point charges are held 1 ft. apart in empty space. What must be the magnitude of each charge in coulombs if they attract each other with a force of 1 lb.?

2. Two equal negative charges, each of magnitude 8 statcoulombs, are held fixed with a separation of 4 cm. Calculate the intensity of the field produced by these charges in a plane which bisects the line joining these charges at right angles, at a point 16 cm. from the intersection of the plane and line joining the charges. At what points of the plane is the intensity a maximum? What is the maximum value of the intensity?

3. Two equal and opposite charges $+q$ and $-q$ are held fixed with a separation $2a$. Using the line connecting the charges as an x-axis with an origin halfway between them, calculate:

a. The intensity of the field produced by this dipole at any point on the xi

b. The intensity of the field at any point on a line perpendicular to the x-axis and passing through the positive charge.

c. The intensity due to each charge separately and from this the resultant intensity (magnitude and direction) at any point in the x-y plane, using a perpendicular bisector of the line joining the charges as a y-axis.

4. Two pith balls, each of mass 0.02 gram, are suspended from a common point by threads of length 10 cm. Each is given the same charge, and in equilibrium the threads make an angle of 74° with each other.

Compute the charge on each pith ball.

5. Charges of $+40$ and -10 statcoulombs are placed along the x-axis at -10 and 0 cm., respectively.

a. Make a plot of the potential as a function of x at any point along the x-axis, also at any point on a line perpendicular to the x-axis through the point $x = +10$ as a function of distance along this line.

b. At what points on the x-axis is the potential 1 statvolt? Is the electric intensity the same at all these points?

c. At what point would a third charge remain in equilibrium? Would it be stable equilibrium?

6. Using Gauss's theorem, prove that the *difference* between the electric intensity vectors on either side of a very large uniform plane sheet of charge (*not* a conductor) is 4π times the surface-charge density. What is the intensity at a point outside the plane?

7. Starting from the fact that for electrostatic equilibrium no electric field can exist in the interior of a metallic conductor, show by the use of Gauss's law that the charge on a charged metal body of arbitrary shape must reside on its surface.

8. Given a charge q distributed uniformly throughout a sphere of radius a with charge density ρ.

a. With the help of Gauss's law show that the electric intensity at points inside the sphere is proportional to the distance from the center of the sphere.

b. Derive a formula for the intensity at points inside and outside the sphere.

c. Make a plot of the magnitude of the electric intensity against distance from the center of the sphere.

d. Calculate the potential as a function of distance from the center of the sphere for points both outside and inside the sphere.

e. Plot the potential as a function of distance from the center of the sphere.

9. A thin wire is bent to form the arc of a circle of radius r, subtending an angle θ_0 at the center of the circle. If the wire carries a charge q uniformly distributed along its length, derive an expression for the intensity of the field at the center of the circle.

10. A spherical drop of water 1 cm. in diameter carries a charge of 5 statcoulombs.

a. What is the potential at the surface of the drop?

b. If two such drops, similarly charged, coalesce to form a single drop, what is the potential at the surface of the drop thus formed?

c. What is the maximum intensity of the field in parts *a* and *b*? Where is the intensity maximum?

11. A conducting sphere of radius a carries a charge q. If an infinitesimal additional charge dq is brought up to the sphere from a distant point, calculate the work done in bringing up this charge. What is the total work

done in charging the sphere to a potential V, considering the sphere initially uncharged and then charged in the manner described above?

12. One face of a thin circular disk of radius R is charged uniformly with a positive surface-charge density σ.

a. Derive an expression for the potential at a point on the axis of the disk at a distance x from the center of the disk.

b. From your answer to part *a* derive an expression for the electric intensity at this point. What is its direction?

c. If the disk is 3 cm. in radius and carries a *total* charge of 0.9 statcoulomb, calculate the potential and intensity at a point on the axis 4 cm. from the center of the disk. Give your answers in volts and volts per centimeter.

d. Using the data of part *c*, calculate the potential at the center of the disk in volts.

e. At what point on the axis is the intensity a maximum? What is this maximum intensity?

13. Using the expression $V = (p \cos \theta)/\epsilon_0 r^2$ for the potential of a simple dipole of moment p at a point whose polar coordinates are r,θ [Eq. (25) of text], calculate:

a. The component of field intensity along r at this point.

b. The component of field intensity perpendicular to r at this point.

c. The resultant intensity and its direction at this point.

14. Calculate the potential at a point outside two neighboring parallel infinite sheets of charge, one carrying a uniform positive surface-charge density $+\sigma$ and the other an equal uniform negative surface-charge density $-\sigma$. The planes have a separation d. Calculate the potential at a point on the other side of the planes, and show that the change in potential as one crosses this so-called double layer is equal to $4\pi\sigma d/\epsilon_0$.

σd is known as the dipole moment per unit area of the double layer.

15. A hemispherical cup of radius r carries a charge q uniformly distributed over its surface. What is the electric intensity at the center of the hemisphere?

CHAPTER III

INDUCED CHARGES AND CAPACITY

Electrostatic problems involving conductors are in general much more complicated than those discussed in the last chapter in which we considered the fields produced by systems of *fixed* charges. The complication arises from the fact (which we have discussed briefly in Chap. I) that, when a conductor is placed in an electrostatic field, induced charges distribute themselves in such a manner as to make the conductor an equipotential. Not only the field but also the distribution of these induced charges on conductors must be calculated, and only for comparatively simple geometrical configurations have exact solutions been obtained.

12. Induced Charges.—If an uncharged insulated conductor is placed in an electrostatic field, the total charge on the conductor stays equal to zero, positive and negative surface charge appearing in equal amounts. Just as many lines terminate on negative induced charges as leave from the positive induced charges. Suppose we consider the case of an uncharged metallic sphere placed in the neighborhood of a positive point charge. Some of the lines of D leaving the point charge will terminate on the side of the sphere toward the point charge, and an equal number will leave from the other side. The magnitude of the induced charge of either sign will depend on how much of the field is intercepted by the conductor. If a charged body is brought inside a hollow metal container (with a small opening), practically all the field is intercepted, and there is an induced charge on the inside of the container equal to, and of sign opposite to, the charge introduced into the container. On the outside of the container one finds a charge equal in sign and magnitude to the charge introduced. If the outside of the container is connected to an electroscope, the leaves will diverge because of this charge. Upon bringing the charged body into contact with the inside of the container, no effect is noted on the electroscope, and upon removing the originally charged body there is still no change in

the electroscope. The charge remaining on the metal container is thus shown to be equal to the original charge introduced. The above experiment was performed along with others by Faraday, using an ice pail as the metal container, and these are known as the Faraday ice-pail experiments. Using this arrangement, Faraday found that, if two bodies were electrified by friction inside the pail, no deflection of the electroscope was observed. If one of the two bodies which were rubbed was removed from the pail, the leaves of the electroscope diverged, thus showing that the bodies were charged. This provides an experimental proof of the law of *conservation of charge*, showing that the two bodies had acquired exactly equal and opposite charges.

In spite of the difficulties of exact calculation, one can obtain a good idea of the field in the presence of conductors by sketching the lines of D and the equipotentials, remembering that the surface of the conductor is an equipotential and that the equipotentials near this surface must bear some resemblance to it. In Fig. 15 is shown the field which results from placing an insulated uncharged sphere in a region where there originally was a uniform field. The solid lines are the lines of displacement (or intensity), and the dotted ones show the equipotentials.

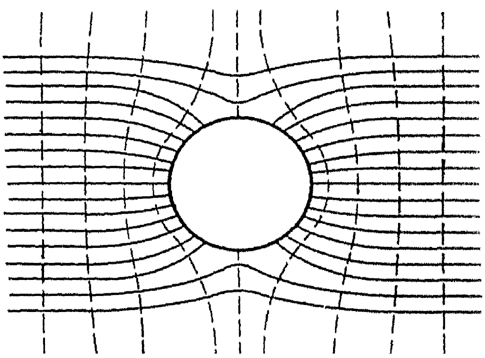

Fig. 15.

13. Method of Electrical Images.—The distribution of charge induced on a conductor when it is placed in an electric field can be calculated in certain simple cases by the method of electrical images, introduced by Lord Kelvin. We shall explain the method with the help of an example. Suppose we have a large plane conductor placed in the field produced by a point charge $+q$. Let the distance d of the point charge q from the plane be small compared to the dimensions of the plane so that we may treat the latter as infinite. Furthermore, the conductor is grounded (connected to the earth) so that we may take its potential as zero. We now must try to find a distribution of charge on the surface of the conductor such that its potential is zero. In the method of images we imagine the conducting plate

removed and a point charge $-q$ placed at a distance d in back of the plane occupied by the conducting surface, as if it were the image of the charge $+q$ in a plane mirror (Fig. 16). The field of these two point charges is such that the plane bisecting the line joining them (the position of the conducting surface) is an equipotential because every point on this plane is equidistant from both charges. Thus at a point P, the potential is

$$V = \frac{q}{\epsilon_0 r} - \frac{q}{\epsilon_0 r} = 0$$

The two point charges produce the same effect as the induced charge and the original point charge. At any point Q we can calculate the field since the potential there is

$$V = \frac{q}{\epsilon_0}\left(\frac{1}{r_1} - \frac{1}{r_2}\right)$$

and this yields the dipole field discussed in Chap. II. In the real problem, we thus have the field at all points to the right of the plane, since there is no field inside or to the left of the plate.

To obtain the distribution of induced charge on the plane, we make use of the fact that the intensity of the electric field at the surface of a conductor equals $4\pi\sigma/\epsilon_0$, where σ is the surface-charge density at the point in question [see Eq. (11), Chap. II]. The intensity $\mathcal{E}$ at an arbitrary point P of the plane is the vector sum of the two vectors $\mathcal{E}_+$ and $\mathcal{E}_-$, as shown in Fig. 16. The vectors $\mathcal{E}_+$ and $\mathcal{E}_-$ are equal in magnitude, each being $q/\epsilon_0 r^2$, and they make equal angles with the x-axis. The resultant is

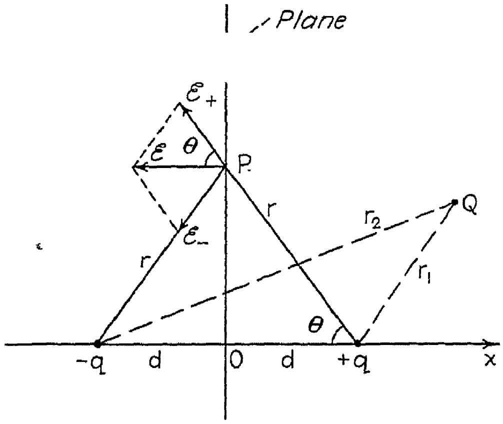

FIG. 16.

$$\mathcal{E} = -\frac{2qd}{\epsilon_0 r^3} \tag{1}$$

the negative sign indicating that it is directed to the left, *i.e.*, toward the surface. The induced surface-charge density at this

point is thus

$$\sigma = \frac{\epsilon_0 \mathcal{E}}{4\pi} = -\frac{qd}{2\pi r^3} \tag{2}$$

and varies inversely as the cube of the distance from the point charge. One can readily show that the total induced charge is just equal to $-q$, equal and opposite to the point charge outside the plane. The proof of this is left to a problem. The force with which the conducting plane and the point charge attract each other is equal to the force with which q attracts its mirror image. Note that, although the conductor is grounded (its potential is zero), it carries a charge. It is a common error to suppose that a grounded conductor can never carry an electric charge.

14. Capacity Coefficients; Condensers. If an isolated conducting body, such as a metal sphere, is charged, its potential (with respect to an infinitely distant point) is raised or lowered depending on the magnitude and sign of the charge imparted to it. In fact, the potential is proportional to the charge on the sphere. If the body is not alone, however, and there are other conducting bodies in its neighborhood, the potential of the first body will depend not only on its own charge but also on the charges and positions of all its neighbors. It can be shown (although we shall not give a proof here) that, for a system of conducting bodies, the potential of each is a *linear* function of the charges carried both by itself and all the others. Conversely, the charge on each conductor is a linear function of the potentials of all the bodies, itself included. The constant coefficients appearing in these relations (charges in terms of potentials) depend only on the geometry of the system, not on its electrical state, and are called the *capacity coefficients* of the system. In our work we shall restrict our attention to the case of *two-body* systems for which the above relations become comparatively simple.

Let us consider two uncharged metallic bodies, insulated from each other and far removed from other conductors or from charged bodies. If we transfer a quantity of charge from one body to the other, as can be done by connecting them for a moment to opposite terminals of a battery or power line, one of the bodies will carry a charge $+q$, the other a charge $-q$, and

there will exist a potential difference between the bodies. Such a two-body system is called an electrical *condenser* or *capacitor*. The potential difference between the two so-called plates of a condenser is proportional to the charge transferred from one plate to the other and *the ratio of the charge on either plate to the potential difference between them is called the electrical capacity or capacitance of the condenser.* If q is the magnitude of the charge on either plate and $V_2 - V_1$ denotes the potential difference between the plates, the capacity C is given by

$$C = \frac{q}{V_2 - V_1} \tag{3}$$

The capacity C of a condenser is defined as a positive quantity, and its value depends only on the geometry of the system and the medium in which the field produced by the charges on its plates exists. We shall, in this discussion, confine our attention to the case of empty space (or air, since the value of ϵ for air is practically the same as for vacuum). In the electrostatic system of units, the dimensions of C can be readily shown to be those of length, so that the unit capacity in this system of units is the *centimeter*, sometimes called a *statfarad*. In the m.k.s. system, the unit capacity, 1 coulomb/volt, is given the name 1 *farad*. One farad is equivalent to 9×10^{11} cm.

All the lines of D which leave the positively charged plate of a condenser must terminate on the negative plate because the charges on the plates are equal and opposite. The general problem of determining this field for arbitrary bodies (in order that the potential difference may be calculated) is virtually impossible, so that capacities are usually determined experimentally. There are, however, several simple and important geometrical configurations for which a theoretical calculation can be made without difficulty and we shall examine a few of these.

a. The Parallel-plate Condenser.—The condenser consists of two parallel metal plates, each of area A, separated by a distance d which is small compared to the surface dimensions of the plates (Fig. 17). If the left-hand plate carries a charge $+q$ and the other plate a charge $-q$, the field due to these charges will exist practically only in the region between the plates. There will be a slight fringing at the ends, but this can be neglected if

the plates are large enough. In the figure are indicated the lines of D originating at the positive charges, which are uniformly distributed on the inside surface of the left-hand plate, and terminating on corresponding negative charges on the opposite plate. The field between the plates is a uniform field, and, from Eq. (10) of Chap. II, we have for the magnitude of D

$$D = 4\pi\sigma \tag{4}$$

By definition, the potential difference between the plates is equal to the work per unit charge which must be done in moving a

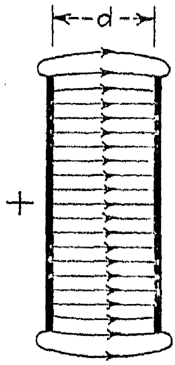

FIG. 17.

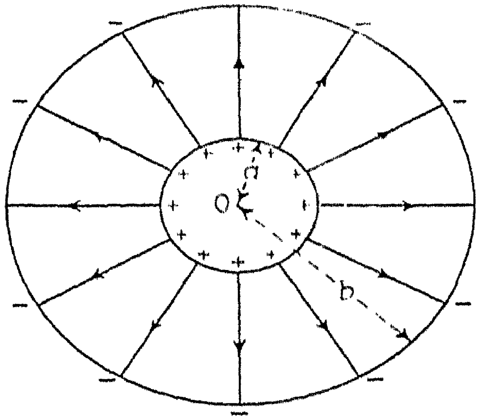

FIG. 18.

charge from one plate to the other. Since the field is uniform, this is

$$\int \mathcal{E}_s \, ds = \mathcal{E}d$$

so that we may write

$$V_2 - V_1 = \mathcal{E}d \tag{5}$$

To calculate the capacity of the condenser, we must have an expression for q, the charge on either plate. From Eq. (4) we find

$$q = A\sigma = \frac{AD}{4\pi}. \tag{6}$$

Using the definition Eq. (3) of capacity, there follows for the capacity of a parallel-plate condenser (neglecting fringing) in vacuum or air,

$$C = \frac{q}{V_2 - V_1} = \frac{D}{\mathcal{E}} \frac{A}{4\pi d} = \epsilon_0 \frac{A}{4\pi d} \tag{7}$$

Equation (7) shows that the capacity increases with increasing area and with decreasing separation of the plates.

b. The Spherical Condenser.—The condenser consists of a metallic sphere of radius a (let us say) and a hollow concentric metallic sphere of inner radius b surrounding the first (Fig. 18). The spherical symmetry of the condenser demands that the charges on the inner and outer spheres be distributed uniformly over the surfaces.

At any point (between the spheres) at a distance r from O, the displacement field points out along r and has a value

$$D = \frac{q}{r^2} \tag{8}$$

with a corresponding intensity in vacuum

$$\mathcal{E} = \frac{q}{\epsilon_0 r^2} \tag{9}$$

since the field is the same as if the charge q were concentrated at the center O. The difference of potential between the plates is

$$V_a - V_b = -\int_b^a \mathcal{E}_r\,dr = -\frac{q}{\epsilon_0}\int_b^a \frac{dr}{r^2} = \frac{q}{\epsilon_0}\left(\frac{1}{a} - \frac{1}{b}\right) \tag{10}$$

and from Eq. (3) the capacity of the condenser is

$$C = \frac{q}{V_a - V_b} = \frac{\epsilon_0}{(1/a) - (1/b)} = \frac{\epsilon_0 ab}{b - a} \tag{11}$$

If the radius b of the outer sphere is very large compared to a, we have

$$C = \epsilon_0 a \tag{12}$$

as the *capacity of an isolated sphere of radius a.* We can give a simple interpretation to the unit of capacitance in the electrostatic system of units. In this system $\epsilon_0 = 1$, and the unit of capacity is the capacity of an isolated sphere of radius 1 cm.

c. The Cylindrical Condenser.—The condenser consists of a metallic cylinder of radius a and a concentric hollow metallic cylinder of inner radius b surrounding the first cylinder. If the length l of the cylinder is very large compared to the separation, the field between the cylinders is essentially that produced by an infinitely long straight wire, uniformly charged. For this case a

calculation similar to those of the preceding examples yields, as the *capacity per unit length* of such a condenser,

$$\frac{C}{l} = \frac{1}{2\ln(b/a)} \tag{13}$$

The proof of this is left as a problem for the reader.

15. Condensers in Parallel and Series. In many practical cases one is interested in the capacity of systems of condensers when they are connected to one another in various manners by metallic wires. At first sight it might seem that we would have to employ the general theory of capacity coefficients to carry through such a calculation. We have seen, however, that the field due to the charges on the plates of a condenser is confined practically entirely to the region between the plates. The presence of other conductors in the neighborhood of a condenser makes no difference, provided only that they are outside the field produced by the condenser. This is the case in practice, and we shall proceed with the calculation on this basis.

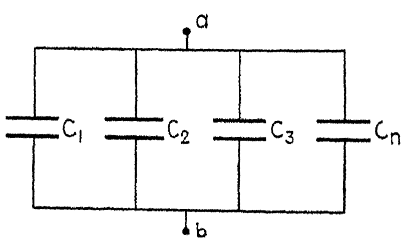

FIG. 19.

Suppose a number n of condensers are connected in parallel, as shown in Fig. 19. The difference of potential V_{ab} is the same for all the condensers $C_1, C_2, \cdots, C_n$ by virtue of the connection, and the total charge on the condensers is the sum of the charges on the individual condensers. For the individual condensers we have

$$q_1 = C_1V_{ab}; \qquad q_2 = C_2V_{ab}; \qquad \cdots; \qquad q_n = C_nV_{ab}$$

If we add all these equations, there follows

$$q_1 + q_2 + \cdots + q_n = Q = V_{ab}(C_1 + C_2 + \cdots + C_n)$$

and, since a single condenser which is equivalent to this combination would take the same charge when its plates have a potential difference V_{ab} between them, its capacity would be

$$C = \frac{Q}{V_{ab}} = C_1 + C_2 + \cdots + C_n \tag{14}$$

Thus the capacity of a single condenser equivalent to a number of condensers in parallel is equal to the sum of the individual capacities.

If we connect the condensers in *series*, we obtain the arrangement shown in Fig. 20. Suppose we establish a definite potential difference V_{ab} between the terminals a and b by connecting them to a battery. Let the positive charge on plate I be $+q$. There must be an equal negative charge $-q$ on plate II since all the lines of D starting from plate I terminate on plate II. Since the plates II and III are connected together and are originally uncharged, there must be an equal charge $+q$ on plate III.

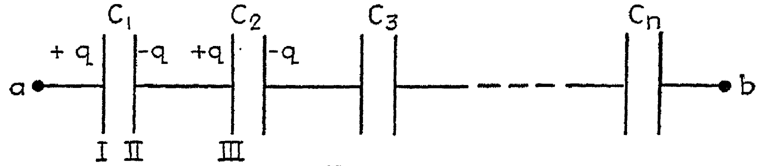

FIG. 20.

Thus we see that the charges on the condensers are equal in series connection. If V_1, V_2, $\cdots$, V_n denote the potential differences between the plates of the various condensers, we see that

$$V_{ab} = V_1 + V_2 + \cdots + V_n$$

since the work one must do to move a charge from b to a is equal to the sum of the works done in moving through the fields of the individual condensers.

For the individual condensers we have (since the charges are all equal)

$$V_1 = \frac{q}{C_1}; \qquad V_2 = \frac{q}{C_2}; \qquad \cdots; \qquad V_n = \frac{q}{C_n}$$

Adding these equations, there follows

$$V_{ab} = \frac{q}{C} = q\left(\frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} + \cdots + \frac{1}{C_n}\right)$$

where C is the capacity of a single equivalent condenser. Hence we have

$$\frac{1}{C} = \frac{1}{C_1} + \frac{1}{C_2} + \cdots + \frac{1}{C_n} \tag{15}$$

as the formula for calculating the capacity of a single condenser which is equivalent to a number of condensers in series.

16. Energy Stored in a Condenser; Energy in the Electrostatic Field.—Let us calculate the work which must be done in charging a condenser. To do so, we imagine the charge brought from one plate to the other in successive steps, an infinitesimal amount dq in each step. Suppose at some point in the process the charge on the plates is q and the potential difference V'. The work we have to do in bringing an additional charge dq from one plate to the other is

$$dW = V'\,dq \tag{16}$$

since V' is the work per unit charge which one must do to transfer a charge from one plate to the other. By definition we have

$$V' = \frac{q}{C} \tag{17}$$

where C is the capacity of the condenser. Substituting in Eq. (16), there follows

$$dW = \frac{1}{C}\,q\,dq$$

and the total work in charging the condenser to a final charge Q is

$$W = \frac{1}{C}\int_0^Q q\,dq = \frac{Q^2}{2C} \tag{18}$$

If the final difference of potential is V, this can be written as

$$W = \tfrac{1}{2}CV^2 \tag{18a}$$

This energy can be thought of as stored up as potential energy in the condenser since it can all be regained. For example, in a parallel-plate condenser one could allow the plates to come together under the action of their attraction for each other and thereby raise a weight. We can use this type of argument to calculate the force of attraction between two condenser plates. Let the condenser plates have a separation x, and let us imagine the condenser charged with charges $\pm Q$ on its plates and insulated so that the charge cannot escape. If we separate the plates by an additional amount dx (Fig. 21), we must do an amount of work

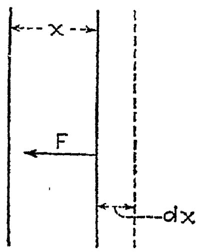

FIG. 21.

$$dW = F\,dx \tag{19}$$

where F is the magnitude of the force of attraction between the plates. This work must be stored up in the condenser. We can calculate the increase of energy stored in the condenser by noting that we have decreased the capacity of the condenser, and hence from Eq. (18) the energy increases. We write for the energy U, in accordance with Eqs. (18) and (7),

$$U = \frac{Q^2}{2C} = \frac{2\pi Q^2}{\epsilon_0 A} x \tag{20}$$

and if x is increased by dx, the increase of U is

$$dU = \frac{2\pi Q^2}{\epsilon_0 A} dx \tag{21}$$

and, equating (21) and (19), there follows for the attractive force

$$F = \frac{2\pi Q^2}{\epsilon_0 A} = \frac{2\pi\sigma^2}{\epsilon_0} A \tag{22}$$

where σ is the surface-charge density on either plate. From this we find for the *force per unit area* (a normal stress) acting on the charged surfaces

$$\frac{F}{A} = \frac{2\pi\sigma^2}{\epsilon_0} \tag{23}$$

This expression turns out to be correct for any charged surface, although we have derived it only for a special case. Note that in the derivation we have considered the charge on the condenser plates constant while the plates are moved. One might ask, "Why not calculate the work done during a motion dx of one of the plates, keeping the potential constant?" Such a calculation would not yield the correct value for the force because one would have to maintain a constant difference of potential with an external device, such as a battery. The battery would then deliver energy to or absorb it from the condenser when the separation of the plates is changed. Hence we could not correctly identify the mechanical work done with the increase of electrostatic energy in the condenser.

Let us return to the question of the energy stored in a parallel-plate condenser. From Eq. (18), this can be written as

$$U = \frac{Q^2}{2C} = \frac{\sigma^2 A^2}{2C} = \frac{4\pi\sigma^2 Ad}{2\epsilon_0} \tag{24}$$

using the fact that the capacity $C = \epsilon_0 A/4\pi d$. It is interesting to write this expression in terms of the field vectors $\mathcal{E}$ and D. We have $D = 4\pi\sigma$ and $\mathcal{E} = 4\pi\sigma/\epsilon_0$ for the uniform field of the parallel-plate condenser, so that Eq. (24) takes the form

$$U = \frac{1}{8\pi}\mathcal{E}D(Ad) = \frac{\epsilon_0\mathcal{E}^2}{8\pi}(Ad) \tag{25}$$

Now Ad is just the volume of space in which the field is different from zero, and the energy is proportional to this volume. Thus for this case we can think of the energy as stored in the electrostatic field with an energy density (energy per unit volume) equal to $\frac{1}{8\pi}\mathcal{E}D$. One can prove in general (using methods which would take us beyond the scope of this book) that the above interpretation is possible for any arbitrary electrostatic field. It takes work to establish such a field, and we can think of this work as being stored up in the field, distributed throughout space with a density given, as above, by

$$u = \frac{1}{8\pi}\mathcal{E}D = \frac{\epsilon_0\mathcal{E}^2}{8\pi} \tag{26}$$

where $\mathcal{E}$ and D represent the magnitudes of the intensity and displacement vectors at the point where the energy density is being calculated. Equation (26) holds only for empty space as written with the factor ϵ_0, but we shall later see that a very similar relation holds in material media. For a given field one can compute the total electrostatic energy by integrating over all space, so that

$$U = \frac{\epsilon_0}{8\pi}\int_{\substack{\text{all}\\\text{space}}} \mathcal{E}^2\, dv \tag{27}$$

where dv is an element of volume.

Problems

1. Prove by direct integration that the total charge induced on an infinite conducting plane by a point charge is equal to and of opposite sign to the point charge. [Use Eq. (2) for the surface density of induced charge.]

2. In the preceding problem, derive an expression for the fraction of the total induced charge on the plane which lies inside a circle whose radius equals the distance of the charge from the plane, the center of the circle

being the point of intersection of a perpendicular from the point charge with the plane.

3. An electron is located at a distance of 10^{-8} cm. in front of a large plane metallic plate.

a. Calculate the force exerted by the plate on the electron when in the above position.

b. When the electron is at a distance x from the plate, calculate the attractive force.

c. How much work (in ergs) would it take to pull the electron to infinity starting at the point $x = 10^{-8}$ cm.?

d. Through what difference of potential in volts would an electron have to move to gain the amount of energy calculated in part *c*?

4. A parallel-plate condenser has plates of 100-cm.^{2} area separated by a distance of 1 mm.

a. Calculate the capacity of this condenser in e.s.u. and in microfarads.

b. If the plates of the condenser are connected to the terminals of a 100-volt battery, what charge resides on the plates?

5. Considering the earth as an isolated spherical conductor, calculate its capacity in microfarads. The radius of the earth is 4,000 miles.

6. Derive Eq. (13) for the capacity per unit length of a long cylindrical condenser.

7. Derive an expression for the capacity of the condenser formed by two spheres, each of radius a, separated on centers by a distance b, b being so much larger than a that the distribution of charge on either sphere may be taken as very nearly unaffected by the presence of the other sphere.

8. Three condensers of capacities 2, 4, and 6 μf, respectively, are connected in series, and a potential difference of 200 volts is established across the whole combination by connecting the free terminals to a battery.

a. Calculate the charge on each condenser.

b. Find the potential difference across each condenser.

c. What is the energy stored in each condenser?

9. The three condensers of the preceding problem are connected in parallel and then connected to the same battery as before.

a. What is the total charge on all three condensers?

b. What is the total energy stored in all three condensers?

10. The 4- and 6-μf condensers of the preceding problems are connected in parallel and the combination then connected in series with the 2-μf condenser. The potential difference across the system so formed is maintained at 200 volts. Calculate the total energy stored in the condensers and the charges on each.

11. A 0.01-μf condenser is alternately charged to a potential difference of 5,000 volts and discharged through a spark gap, 500 times per second. What is the energy dissipated per discharge? What is the average rate of power dissipation?

12. Suppose two condensers, one charged and the other uncharged, are connected together in parallel. Prove that, when equilibrium is reached, each condenser carries a fraction of the initial charge equal to the ratio of its capacity to the sum of the two capacities. Show that the final energy

in the system is always less than the initial energy and derive a formula for this difference in terms of the initial charge and the capacities of the condensers.

13. Two condensers of capacities 5 and 10 μf are each charged with a potential difference of 100 volts and the negative plate of the 5-μf condenser is connected to the positive plate of the 10-μf condenser.

a. What is the total charge on the two plates which are connected together?

b. If the other two plates are then connected together, what is the potential difference across each condenser after equilibrium is established?

14. Show that the force of attraction per unit area between two parallel condenser plates can be written as $\frac{F}{A} = \frac{\epsilon_0}{8\pi}\mathcal{E}^2$, where $\mathcal{E}$ is the intensity of the field between the plates. From this show that, if the potential difference between the plates is maintained constant (by connecting the condenser to a battery), the force varies inversely as the square of the separation of the plates.

15. A parallel-plate condenser having plate areas of 100 cm.^{2} and a separation of 2 mm. is permanently connected to a 100-volt battery. Using the results of the preceding problem, calculate how much work is done in separating the plates to a distance of 4 mm., maintaining the potential difference constant during the process. Does the energy stored in the condenser increase or decrease, and by how much? Compare this with the mechanical work done in effecting the plate separation.

16. Assuming the validity of the expression $2\pi\sigma^2/\epsilon_0$ as the force per unit area acting outward on any charged conducting surface (σ the surface-charge density), calculate the maximum charge which can be put on the surface of a water drop 2 cm. in diameter. The surface tension of water is 72 dynes/cm. What is the potential at the surface of the drop under these conditions?

HINT: The maximum charge is that for which the outward electrical force just balances the surface-tension force.

17. For the problem of the point charge and conducting plane, calculate the force of attraction between charge and plane by integrating the expression $2\pi\sigma^2/\epsilon_0$ over the area of the plane. Show that this is just the same as the attraction of the point charge for its image.

18. Show that the expression $Q^2/2C$ gives the energy stored in a spherical condenser by integrating the energy density $u = \frac{\epsilon_0}{8\pi}\mathcal{E}^2$ over the region of space between the plates. Take as volume element the volume between two spheres of radii r and $r + dr$, respectively.

19. Assuming that an electron is a uniformly charged sphere of radius a, calculate an expression for the total electrostatic energy in the field produced by a single electron.

Assuming further that this energy is equal to mc^2, m the mass of an electron, c the velocity of light (3×10^{10} cm./sec.), calculate the radius of an electron in centimeters.

CHAPTER IV

STEADY ELECTRIC CURRENTS

The fundamental laws of the equilibrium behavior of conductors in electrostatic fields are changed radically when equilibrium is disturbed, *i.e.*, when there is "flow" or motion of electric charges. When the charges in a conductor are not at rest, the intensity of the field at the surface need not be normal to the surface. Hence the conducting body is not an equipotential, and there will exist an electric field inside the conductor. As an example let us consider a charged condenser in electrostatic equilibrium. If we connect the two plates by a metallic wire, we have, at the instant of contact, a single conducting system (the two plates and the wire) with a difference of potential between two of its parts. Since this is not a possible equilibrium state, electric charge flows from one plate to the other until all points of the system attain the same potential. We say that an *electric current* flows in the wire connecting the plates and in this example the flow is *transient* or nonsteady, equilibrium being reestablished in a very short time. One can say that positive charge flows from the plate of higher potential to that of lower potential, or that negative charge flows in the opposite direction, or both. No matter what picture one has, it is necessary to adopt a convention as to the direction of current flow, and we do so by calling *the direction of flow of positive charge the direction of flow of the electric current.* When electric current flows in a material medium, such as a conductor, it is usual to denote the current as *conduction current,* whereas, if charged masses such as electrons or ions transport the charge directly from one point to another, the current is called *convection current.* The essential point is that charge is transferred from one point to another, the whole situation being analogous to the mode of description of heat flow. In fact, the analogy is so close that the student should constantly draw parallels between the two phenomena.

17. Definitions of Current and Current Density.—We shall define the electric current flowing across a definite surface as

the charge per unit time crossing this surface. In symbols

$$i = \frac{dq}{dt} \tag{1}$$

where dq is the charge passing across the surface in time dt. For many problems it becomes necessary to introduce the idea of *current density*. Consider an infinitesimal area at a point of a medium carrying current, and let us choose the elementary area so that it is perpendicular to the direction of current flow. If we denote this area by dS_n and the current crossing it by di, then the *current density* j at this point is defined by

$$j = \frac{di}{dS_n} \tag{2}$$

The subscript n is to indicate that dS_n is normal to the direction of current. If j is known (magnitude and direction) at every

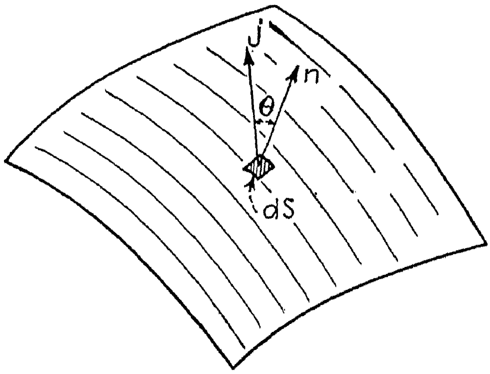

FIG. 22.

point of a surface, we can obtain the total current crossing a finite area by the following scheme: The current crossing an element dS of the area is $j_n\, dS$, since only the component of j normal to dS contributes to the charge crossing this area. The total current is then the sum of the contributions from all the elements of the area under consideration. Thus we have (Fig. 22)

$$i = \int_{\text{area}} j_n\, dS = \int j \cos\theta\, dS \tag{3}$$

In the electrostatic system of units, the unit of current is 1 statcoulomb/sec. and is called a *statampere*. The unit of current

density is thus 1 statampere/cm.[2] In the m.k.s. system the unit current of 1 coulomb/sec. is called 1 *ampere* and the corresponding unit of current density is 1 *ampere per square meter.*

18. The Steady State: Equation of Continuity.—Consider a conducting medium in which electric current is flowing, and let us imagine that at each point we construct the vector j. The totality of such vectors forms a vector field of flow just as in the case of flow of fluids. One can construct "lines of flow" which give the direction of the current at each point of space and also tubes of flow just as in the hydrodynamical case. In this chapter we restrict our attention to the case of the steady or stationary state in which the pattern of lines of flow stays fixed. The current density maintains a definite value at each point of the medium although it may vary from point to point. In

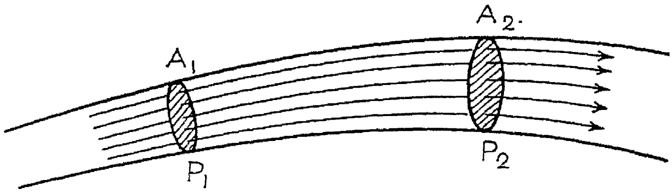

FIG. 23.

the applications with which we shall concern ourselves, we will have to do largely with the flow of currents in metallic wires. In this case we have a bundle of stream lines within the wire (the current flowing along the length of the wire) which we consider to form a single tube of flow.

The law of conservation of electric charge places a definite restriction on the pattern of flow lines, and this is called the *equation of continuity.* We shall derive this law for the steady state. Consider a single tube of flow, as shown in Fig. 23, and let the cross sections at points P_1 and P_2 normal to the direction of current flow be A_1 and A_2, as shown. We shall assume that the current density j is constant at all points of either area. This involves no loss of generality, since we can make these areas as small as we please. In the steady state, the charge entering the volume of the tube between the areas A_1 and A_2 across A_1 per unit time must be just equal to the charge leaving per unit time across A_2. Otherwise the charge in this volume would keep increasing (or decreasing) indefinitely, contradicting the assumption of a steady state. From the law of conservation of

charge we know that no net charge can be created or destroyed inside this volume.

The current across A_1 is $i_1 = j_1A_1$ [Eq. (3)], and across A_2 it is $i_2 = A_2j_2$, so that the equation of continuity takes the form

$$i_1 = i_2$$

or

$$j_1A_1 = j_2A_2 \tag{4}$$

where 1 and 2 refer to *any* pair of points on the tube of flow.

This equation may readily be extended to the case where a tube of flow branches into two or more tubes of flow, as in the case of linear metallic circuits where more than two wires are con-

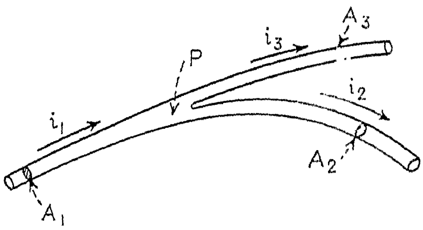

FIG. 24.

nected at a single point (a so-called *branch point*). In Fig. 24 we show the case of three conductors connected at a branch point P. Choosing areas A_1, A_2, and A_3 as shown, one finds easily that the equation of continuity becomes

$$i_1 = i_2 + i_3$$

or

$$j_1A_1 = j_2A_2 + j_3A_3 \tag{5}$$

The extension to the case of more than three conductors meeting at a point is obvious.

The law of conservation of charge requires that, in the steady state, the lines of current flow (and hence the tubes of flow) form *closed* curves and cannot start or stop anywhere in the medium where current is flowing. This fact can be easily formulated mathematically to give a somewhat more general form of the equation of continuity than we have written. Consider any *closed* surface. From the above we see that there can

be no net current entering or leaving the volume enclosed by this surface, and hence we can write

$$\underset{\substack{\text{closed}\\ \text{surface}}}{} dS = 0 \tag{6}$$

This is the general form for the equation of continuity for the steady state. Equations (4) and (5) can be obtained immediately by applying Eq. (6) to the particular cases described by them.

19. Sources of E.m.f.—Thus far we have considered only the geometrical description of the flow of electric currents, and now we must investigate the methods by which such flow can be set up and maintained. Let us start with the simplest case of a single metallic wire. If we maintain a constant difference of potential between the ends of the wire, it is found that a steady state is established in which a constant current flows in the wire and a steady evolution of heat is observed. This generation of heat corresponds to a continual dissipation of energy, and this energy must be supplied by the device which maintains the difference of potential between the ends of the wire. *Any device which maintains a definite difference of potential between two points,* which we call the *terminals,* shall be called a *seat of electromotive force.* If a wire is connected between the terminals of a seat of e.m.f., a steady current will be set up and a definite potential difference will be maintained between the terminals. This potential difference will in general be different for different currents, but in any given case it will be constant, and the seat of e.m.f. will continually supply energy to the electrical circuit thus formed, forcing charge internally from its low potential to its high potential terminal. One important characteristic of steady currents must be always kept in mind: *Although the charges are not at rest in the system, there is a static distribution of potential and of electric field which is maintained by the seat of e.m.f.*

Let us examine the situation in the case of a typical seat of e.m.f., a battery. First, let us suppose that we have a charged parallel-plate condenser with the plates made of copper. If the charged plates are dipped into a dilute sulphuric acid solution (a conducting medium), the electric field between the plates sets the charged ions in the solution into motion; a current

flows discharging the condenser. When equilibrium is reached, the plates are at the same potential. Were we to perform the above experiment with the condenser initially uncharged, nothing would happen when the plates were dipped into the solution. If we substitute a zinc plate for one of the copper plates and dip the uncharged condenser into the sulphuric acid solution, we find the surprising result that a difference of potential appears and is maintained between the plates, the copper being at a higher potential than the zinc. The electric field inside this so-called cell tends to send current from the copper to the zinc and destroy the difference of potential. Since, however, this does not occur, we must assume the existence of forces inside the cell which are not due to the electric field and which we shall denote as *non-electrical* or *chemical* forces. In equilibrium, then, the chemical force tending to drive an ion from one plate to the other is just equal and opposite to the force on the ion due to the presence of the electric field. If we imagine that we were to move an ion inside the cell from the negative plate to the positive plate, the work done by the electric field must be just equal and opposite to the work done by the chemical forces. Thus the work done per unit charge by the chemical forces in the above motion is just equal to the potential difference between the terminals (this is only true on open circuit, *i.e.*, when no current is being sent around a circuit containing the cell) and is called the *electromotive force* of the *cell*. It is in this manner that we can utilize the idea of electromotive force to measure the work per unit charge done by nonelectrical forces which must act in any seat of e.m.f.

There is another method of describing the above situation which is often employed, in which one uses the terminology e.m.f. to denote the work done per unit charge by both non-electrical and electrical forces on a charge carried around a closed path, part of which lies inside the seat of e.m.f. Although there is no essential difference between the two modes of description, it often causes confusion. Let us imagine that we have a cell on open circuit and that we carry a charge around a closed path, as shown in Fig. 25. The work done per unit charge in the part of the closed path AB external to the cell is just V_{AB}, the difference of potential between the plates. Inside the cell we have equal and opposite electrical and nonelectrical forces,

so that the net work done by all the forces for the path BA inside the cell is zero. In either mode of description we have the fundamental result that *the electromotive force of a seat of e.m.f. is measured by the potential difference between its terminals on open circuit.* In symbols we write

$$E = V_{AB} \qquad \text{(open circuit)} \qquad (7)$$

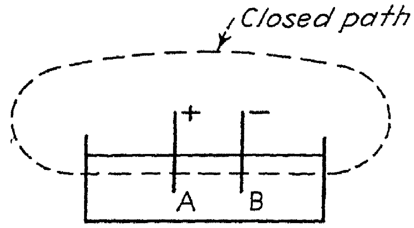

FIG. 25.

One more fundamental point must be insisted upon. Since we have a static electric field maintained by virtue of the action of the seat of e.m.f., we can still write the fundamental law of electrostatics for $\mathcal{E}(= -\text{grad } V)$

$$\oint_{\substack{\text{closed}\\ \text{path}}} \mathcal{E}_s \, ds = 0 \qquad (8)$$

or, in words, the work done by the electrical forces per unit charge around any *closed* path must be zero. This is just another way of stating that the field is conservative and that a potential exists.

20. Ohm's Law for Linear Conductors.—Let us return to the case of a simple series circuit, a single metallic wire of uniform cross section connected to the terminals of a seat of e.m.f. Since a constant potential difference is maintained between the ends of the wire, there must be an electrostatic field within the wire. Under the influence of this field a steady unidirectional current i is maintained. If we change the potential difference between the ends of the conductor, the steady current also changes, and it turns out that the ratio of potential difference to current is *constant,* independent of the current, provided the temperature of the conductor is maintained at a constant value. This is the essence of *Ohm's law.* Denoting the potential difference by V_{ab}, a and b referring to the ends of the wire, then Ohm's law states that

$$\frac{V_{ab}}{i} = \text{constant} \qquad (9)$$

This constant ratio is called the *electrical resistance* of the conductor for this type of current flow. Note that Ohm's law is a statement of the behavior of conducting bodies and in this sense

should be looked upon as describing a property of matter rather than as a fundamental electrical principle. If we denote this constant resistance by R_{ab}, Eq. (9) becomes

$$\frac{V_{ab}}{i} = R_{ab} \tag{10}$$

The unit of resistance in e.s.u. is 1 statvolt/statampere and is called a *statohm*. In the m.k.s. system the unit resistance is 1 volt/amp. and is called 1 *ohm*. Ohm's law as expressed by Eq. (9) or (10) holds for any portion of a linear conductor, a and b then referring to any two cross sections of the uniform wire. For a homogeneous straight wire of uniform cross section, the equipotential surfaces are cross-sectional areas of the wire normal to the direction of the lines of current flow, and the electric field inside the wire is uniform and directed along the lines of flow. We return to this point in the next section.

Since heat is evolved as long as current flows, we think of the expression iR_{ab} as the work done per unit charge on the moving charges by forces which have the nature of friction forces and are directed opposite to the direction of the current. Thus we can interpret the expression iR_{ab} as the work done per unit charge *against* the resistance forces in moving charges from a and b, and it is often called the iR drop along the wire.

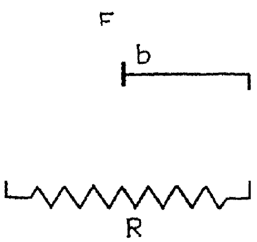

FIG. 26.

A schematic diagram of a simple series circuit is shown in Fig. 26. Let the e.m.f. of the battery be E and let us suppose that the internal resistance of the battery (the resistance of the path carrying current between the terminals) is so small compared to the resistance R of the wire that we can neglect it. The current i flows in the external circuit from the positive to the negative terminal. Let us calculate the work done by the electric field in moving a charge completely around the closed circuit. From Eq. (8) this must be zero. Thus we have

$$\underset{\text{(external)}}{V_{ab}} + \underset{\text{(internal)}}{V_{ba}} = 0 \tag{11}$$

where the first term is taken along the external circuit and the second inside the seat of e.m.f. From Eq. (10) the first term

is iR and from Eq. (7) we see that the second term is $-E$ (using the facts that inside the cell

$$V_{ba} = V_b - V_a = -(V_a - V_b) = -V_{ab}$$

and that there is no internal resistance). Thus Eq. (11) becomes

$$iR - E = 0$$

or

$$E = iR \tag{12}$$

Now let us drop the restriction of no internal resistance, and let us assume that there is an ohmic resistance R_B between the terminals of the battery. When a current flows through a battery, the e.m.f. no longer equals the potential difference between its terminals. If current flows inside the battery from the negative to the positive terminal (the case of Fig. 22), then some of the work done per unit charge by the *nonelectrical* forces in driving charge through the battery is used in overcoming the friction forces (ohmic resistance), and the remaining work is available to do work against the electrical forces. Thus the potential difference between the terminals is less than that on open circuit, and we have

$$V_{ab} = E - iR_B \tag{13}$$

The terminal potential difference is thus the e.m.f. minus the iR drop through the battery.

If current is forced through the battery from the positive to the negative terminal (this can only be done with the help of additional seats of e.m.f.), then the electric forces inside the battery must be larger than the chemical forces. Indeed the work done by the electrical field in moving a charge in this manner between the terminals must be equal and opposite to the sum of the works done by the chemical and friction forces acting on the charge. Expressing this per unit charge, we have in contrast to Eq. (13)

$$V_{ab} = E + iR_B \tag{14}$$

Equation (14) holds, for example, in the case of the charging of a storage battery, whereas Eq. (13) holds when the same battery is discharging.

Equation (12) is no longer correct for the circuit of Fig. 22 when the internal resistance R_B is included, but Eq. (11) is of course still valid. The second term in Eq. (11), referring to the portion of the path inside the battery, is again

$$V_{ba} = -V_{ab} = -E + iR_B$$

using Eq. (13). The first term is iR, so that we obtain

$$iR - E + iR_B = 0$$

or

$$E = i(R + R_B) \tag{15}$$

as the relation between e.m.f., current, and resistances which is to replace Eq. (12). Note that Eqs. (12) and (15) become identical when $R_B = 0$. We see from the foregoing example that Ohm's law in the form of Eq. (10) holds only for a portion of circuit in which there are no seats of e.m.f.

Equation (15) no longer contains any term referring to the work done by purely electrostatic forces. This is due to the fact that it refers to work done in a *closed* path, and according to Eq. (8) the electrostatic forces yield no contribution. The electromotive force of the battery (E), which is the work done per unit charge in moving a charge through the cell, is equal to the work done per unit charge against the dissipative forces in the conducting path.

Resistances in Series and in Parallel.—Let us suppose we have a number n of resistances connected in series as shown in Fig. 27.

FIG. 27.

From the equation of continuity we see that the current flowing in any resistance is the same as that in any other. Thus we have a common current i in a series circuit. If the potential drops across the individual resistances are $V_1, V_2, \cdots, V_n$, Ohm's law yields

$$V_1 = iR_1; \qquad V_2 = iR_2; \qquad \cdots; \qquad V_n = iR_n$$

and, since the potential difference V_{ab} between the outside terminals a and b is equal to the sum of the potential differences

across the individual resistances, there follows

$$V_{ab} = V_1 + V_2 + \cdots + V_n = i(R_1 + R_2 + \cdots + R_n)$$

Thus the series combination of resistances behaves just like a single resistance R, where

$$R = R_1 + R_2 + \cdots + R_n \quad (16)$$

If the resistances are connected in parallel, we have the arrangement shown in Fig. 28. If a constant potential difference V_{ab} is maintained between terminals ab, the potential difference between the ends of any one of the resistances is equal to V_{ab} by virtue of the connection. From Ohm's law

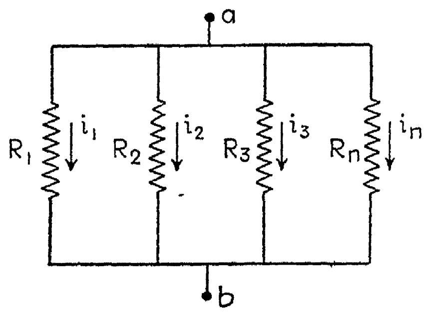

FIG. 28.

$$V_{ab} = i_1R_1 = i_2R_2 = \cdots = i_nR_n$$

or

$$i_1 = \frac{V_{ab}}{R_1}; \qquad i_2 = \frac{V_{ab}}{R_2}; \qquad \cdots; \qquad i_n = \frac{V_{ab}}{R_n}$$

The equation of continuity requires that the total current i entering at a be equal to the sum of the currents $i_1 \cdots i_n$.

$$i = i_1 + i_2 + \cdots i_n = V_{ab}\left(\frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_n}\right)$$

Thus the parallel combination of resistances behaves as a single resistance R, where

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_n} \tag{17}$$

This shows that the equivalent resistance of a parallel combination is always smaller than the smallest resistance in the combination.

Examples

1. Two batteries of e.m.fs. 6 and 8 volts are connected in series as shown in Fig. 29, and a parallel-series combination of resistances is connected to the battery terminals as shown. The internal resistance of the 6-volt battery is 0.4 ohm and that of the 8-volt battery is 0.6 ohm. Required are the currents flowing through the individual resistances, and the potential differences between the battery terminals.

Let the currents be i, i_1, and i_2, as indicated.

From the equation of continuity we have at the branch point a,

$$i = i_1 + i_2$$

We can reduce the problem to that of a simple series circuit by replacing the parallel combination R_1 and R_2 by an equivalent resistance R', where

$$\frac{1}{R'} = \frac{1}{R_1} + \frac{1}{R_2}$$

It is evident that the two batteries have an e.m.f. together equal to the sum of the e.m.fs. of each, and hence Eq. (15) becomes

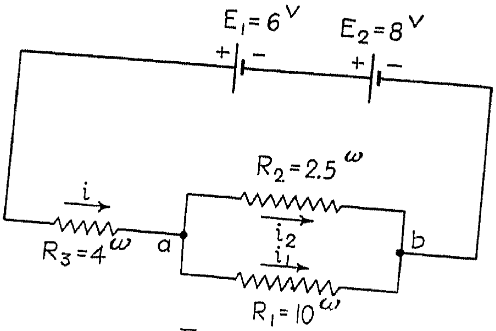

FIG. 29.

$$E_1 + E_2 = i(R' + R_3 + R_{B_1} + R_{B_2})$$

or

$$i = \frac{E_1 + E_2}{R' + R_3 + R_{B_1} + R_{B_2}}$$

For R' we have

$$\frac{1}{R'} = \frac{1}{10} + \frac{1}{2.5} = \frac{5}{10};$$

$$R' = 2 \text{ ohms}$$

so that

$$i = \frac{6 + 8}{2 + 4 + 0.4 + 0.6} = \frac{14}{7} = 2 \text{ amp.}$$

To find the currents in R_1 and R_2, we can write

$$V_{ab} = i_1 R_1 = i_2 R_2$$

where V_{ab} is the potential drop between points a and b. This gives

$$\frac{i_1}{i_2} = \frac{R_2}{R_1} = \frac{2.5}{10} = \frac{1}{4}$$

This, together with the equation of continuity $i = 2 = i_1 + i_2$, gives

$$i_1 = 0.4 \text{ amp.}; \qquad i_2 = 1.6 \text{ amp.}$$

To find the potential difference across the battery terminals, we may utilize Eq. (13) since both batteries are discharging. For the 6-volt battery

$$V_6 = E_1 - iR_{B_1} = 6 - 2 \times 0.4 = 5.2 \text{ volts}$$

and for the other

$$V_8 = E_2 - iR_{B_2} = 8 - 2 \times 0.6 = 6.8 \text{ volts}$$

This problem may also be solved by applying Eq. (8) which may be stated in words as follows: The algebraic sum of the potential drops around any closed loop is zero. In applying this rule, the potential drops iR are positive if one moves *in* the direction of current flow, and potential rises are treated as negative potential drops. The student should work the above problem by this method. We shall return to a detailed consideration of this scheme in a later section.

2. As a second example, let us consider the problem of determining the value of a resistance by the so-called ammeter-voltmeter method. For our purposes it is only necessary to state that an ammeter connected in a circuit reads the current flowing through it and acts only as a small resistance. The voltmeter is a similar instrument which, when connected between any two points of a circuit, reads the potential difference between these points. For reasons which will become evident from our problem, voltmeters have relatively high resistances.

One possible method of connection is shown in Fig. 30. The terminals a and b are maintained at a definite potential difference V_{ab} by some sort of battery or d.c. generator. The voltmeter of resistance R_v is connected directly across the terminals of the unknown resistance R, and the ammeter A is connected in series with the combination. Let the currents flowing through ammeter, resistance, and voltmeter be i, i_r and i_v, respectively, as shown. The reading of the ammeter gives i; that of the voltmeter gives V_{12}, the potential drop across the resistor; and let us suppose that R_v is known.

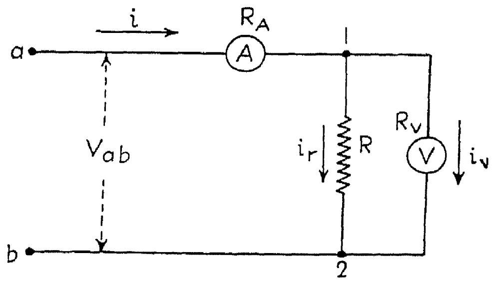

FIG. 30.

At the branch point 1 we have

$$i = i_r + i_v$$

and from Ohm's law

$$V_{12} = i_r R = i_v R_v$$

or

$$i_r = \frac{V_{12}}{R}; \qquad i_v = \frac{V_{12}}{R_v}$$

so that

$$i = V_{12}\left(\frac{1}{R} + \frac{1}{R_v}\right) \tag{18}$$

from which R may be readily found.

If we write the above equation in the form

$$R = \frac{V_{12}}{i}\left(1 + \frac{R}{R_v}\right)$$

we see that, if the voltmeter resistance R_v is very large compared to the unknown resistance R, one may obtain the value of R very nearly by writing V_{12}/i. If, however, R_v is not large compared to R, one must use the complete formula and thus correct for the presence of the measuring instrument.

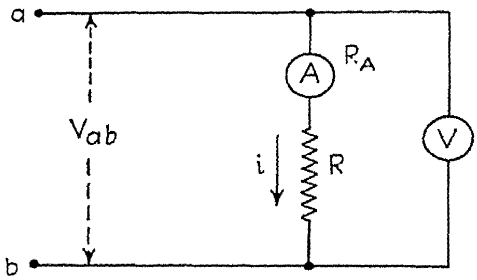

FIG. 31.

One might suppose that the difficulty might be avoided by connecting the elements as shown in Fig. 31. This does not eliminate the disturbing effect of the instruments. Let the current through the resistance R and ammeter be i as shown and the ammeter resistance be R_A. The voltmeter reading is just V_{ab}, since its terminals are connected to the points a and b. Since the potential drop across ammeter and resistance in series is also V_{ab}, we have from Ohm's law

$$V_{ab} = i(R + R_A)$$

or

$$R = \frac{V_{ab} - iR_A}{i} \tag{19}$$

as the correct formula for R, requiring a knowledge of the ammeter resistance. If we write this relation in the form

$$R = \frac{V_{ab}}{i\left(1 + \dfrac{R_A}{R}\right)}$$

we see that again an approximate value of R may be obtained from V_{ab}/i *provided* $R_A \ll R$. Ammeters are constructed so as to have very small resistances in order that instrument corrections may be neglected in most practical work.

21. Resistivity and Conductivity; Ohm's Law for Extended Media.—In order to extend the considerations of the last section to the case of current flow in extended media, it is necessary to examine Ohm's law in some detail with the object of formulating it in such a manner that it can be applied at any point of a conducting medium rather than to a section of a linear conductor.

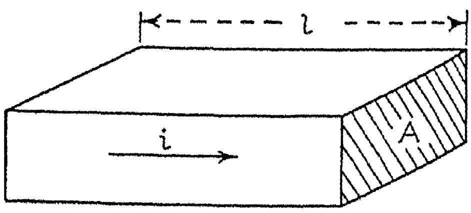

FIG. 32.

As a preliminary step, let us consider a length l of a wire of uniform cross section A (Fig. 32). It is an experimental fact that the resistance of such a conductor between its end

faces is proportional to the length l and inversely proportional to the cross section A. Writing this as an equation, we have

$$R = \frac{\rho l}{A} = \frac{1}{\sigma} \frac{l}{A} \tag{20}$$

where the proportionality constants ρ and σ are known as the *resistivity* and *conductivity* of the material of which the conductor is composed, and obviously we have $\rho = 1/\sigma$.

The unit of resistivity in the electrostatic system of units is 1 statohm-cm., and the student can easily show that this is equal to 1 sec., so that the dimensions of resistivity in e.s.u. are those of time. In the m.k.s. system the unit resistivity is, strictly speaking, 1 ohm-meter. In engineering practice, however, there has grown a custom of specifying resistivity as "ohms per mil-foot," and this needs explanation. This mixed unit has come from the practice of specifying the lengths of wires in feet and the cross section in so-called "circular mils." One mil is $\frac{1}{1000}$ inch and 1 circular mil is the area of a circle of diameter equal to 1 mil. Thus the cross-sectional area of a wire of diameter d mils is equal to d^2 circular mils. In using the "ohm per mil-foot" unit of resistivity, one must be careful in employing Eq. (20) to specify l in feet and A in circular mils to obtain the resistance R in ohms.

The resistivity of copper is about 10 "ohms per mil-foot," 2×10^{-8} ohm-meter, or about 2×10^{-18} e.s.u. In changing from one system of units to another a convenient relation to remember is 1 statohm $= 9 \times 10^{11}$ ohms. The resistivity of a metal varies markedly with temperature, increasing with increasing temperature. For moderate temperature ranges the resistivity can be represented by a linear function of temperature (in a manner exactly like the expansion of a solid)

$$\rho = \rho_0(1 + \alpha t) \tag{21}$$

where α, the temperature coefficient of resistivity (referred to 0°C.), is of the order of magnitude of 0.5 per cent per degree centigrade for ordinary metals. ρ_0 is the resistivity at 0°C.

Returning to our original problem, we now apply Ohm's law to an infinitesimal volume element inside a conducting medium. At an arbitrary point O we construct the tiny cube $dx\,dy\,dz$, as

shown in Fig. 33. If there is no seat of e.m.f. at the point where the volume element is located, Ohm's law in the form of Eq. (10) yields

$$V_{x=0} - V_{x=dx} = -dV = iR \tag{22}$$

where $-dV$ is the negative increase (the drop) in potential between the left- and right-hand faces of the volume element, i is the current flowing normally to the faces $dy\,dz$, and R is the resistance between these faces. From Eq. (3) we have $i = j_x\,dy\,dz$, and from Eq. (20) $R = \dfrac{dx}{\sigma\,dy\,dz}$. Substituting these values in Eq. (22), there follows

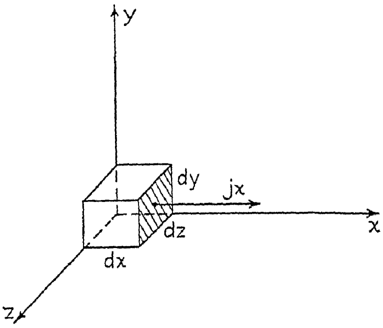

FIG. 33.

$$-dV = j_x\,dy\,dz\frac{dx}{\sigma\,dy\,dz} = \frac{j_x}{\sigma}\,dx$$

Now $-(dV/dx) = \mathcal{E}_x$, the negative x-component of the gradient of the potential, so that

$$\mathcal{E}_x = \frac{j_x}{\sigma}$$

or

$$j_x = \sigma\mathcal{E}_x \tag{23}$$

In words, the x-component of the current density at a point of a conducting medium is equal to the conductivity of the medium times the x-component of the intensity of the field at that point. It is clear that Eq. (23) holds for any component (x, y, or z), and hence, if the medium is isotropic, we can write a single *vector* equation

$$\vec{j} = \sigma\vec{\mathcal{E}} \tag{24}$$

This is Ohm's law in a form which holds at each point of the medium. Equation (24) implies that, for isotropic media obeying Ohm's law, the direction of current flow (for steady currents) coincides with the direction of the electric field intensity at every point. Thus the lines of current flow *coincide* with the lines of electric intensity and are perpendicular to the equi-

potential surfaces. This is by no means an evident fact as one can see by considering the steady flow of current between two electrodes in vacuum where the current is carried by moving electrons. Here the electrons obey Newton's laws of motion. In general the motion of a body is *not* in the direction of the resultant force acting on it. If the lines of $\mathcal{E}$ are not straight lines, then the electron trajectories will certainly not coincide with the lines of $\mathcal{E}$.

22. Kirchhoff's Rules.—Problems involving the steady flow of currents (direct currents) in networks of linear conductors can be solved by a straightforward application of the methods developed in Secs. 18 to 20, but in complex cases it is convenient to

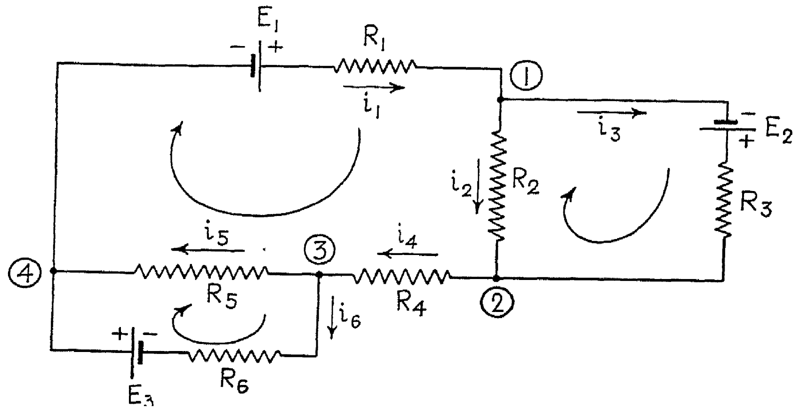

FIG. 34.

follow a systematic procedure which is embodied in *Kirchhoff's rules.* At the outset we wish to emphasize the fact that these rules provide no new principle beyond those already presented. By a network one means a system of linear conductors and seats of e.m.f. interconnected in some arbitrary fashion. A typical network is shown in Fig. 34. In this network there are four *branch points* labeled ①, ②, ③, and ④. Kirchhoff's first rule states: *At any branch point the sum of the currents entering equals the sum of the currents leaving the junction.* This is simply a statement of the equation of continuity as we have formulated it in Eqs. (4) and (5). For example, at branch point ① we have the equation

$$i_1 = i_2 + i_3$$

If there are n branch points, there will be $n - 1$ independent current relations of the type given above.

Kirchhoff's second rule is a statement of the content of Eq. (8). *The sum of the potential drops around any closed loop of the network equals zero.* In applying this rule one must remember that the potential drop across a resistance is positive if one moves *in* the direction of current flow and that potential rises are to be handled as negative drops of potential. Thus in the case of the network of Fig. 34, if we apply this rule to the loop (or mesh) containing E_1, R_1, R_2, R_4, and R_5, we have (starting at point ④ and proceeding clockwise)

$$-E_1 + R_1i_1 + R_2i_2 + R_4i_4 + R_5i_5 = 0$$

The directions of the currents in the various branches are assumed, and negative answers for the currents indicate that the corresponding directions must be reversed. In applying Kirchhoff's rules it is necessary to write down a number of *independent* equations equal to the number of unknowns. It is a common failing to formulate a perfectly correct equation which may be obtained, let us say, by adding two equations already formulated. Clearly this yields no more information than the first two, and it becomes necessary to adopt a systematic mode of procedure. We can describe this best by an example. In Fig. 34 there are four branch points and the equation of continuity demands that

$$\left.\begin{array}{lr} ① & i_1 = i_2 + i_3 \\ ② & i_2 + i_3 = i_4 \\ ③ & i_4 = i_5 + i_6 \\ ④ & i_5 + i_6 = i_1 \end{array}\right\} \qquad (25)$$

The last equation is not independent of the preceding three, and in fact one can obtain it by adding the first three equations. Thus there are *three* independent current relations. Since there are six unknown currents (assuming that the E's and R's are known), we need three more equations to effect a solution. Let us start with the mesh equation already written

$$-E_1 + R_1i_1 + R_2i_2 + R_4i_4 + R_5i_5 = 0 \qquad (26)$$

If now we choose a second closed path not containing E_1, for example, then we are sure that the resulting equation will be

independent of Eq. (26). Thus in the mesh containing E_2, R_2, and R_3, we have

$$-E_2 + i_3R_3 - i_2R_2 = 0 \tag{27}$$

Finally the mesh containing E_3, R_5, and R_6 contains neither E_1 nor E_2; hence the equation relating to it must be independent of the two already written. There follows

$$i_6R_6 - E_3 - i_5R_5 = 0 \tag{28}$$

This completes the task of finding three independent loop equations and illustrates the procedure to be followed. Any other equation, as for example the one obtained by proceeding through E_1, R_1, R_2, R_4, R_6, E_3, and back to E_1, yields

$$-E_1 + i_1R_1 + i_2R_2 + i_4R_4 + i_6R_6 - E_3 = 0$$

and this is simply the sum of Eqs. (26) and (28).

Examples

1. Consider the network of Fig. 35 with $E_1 = 6$ volts, $E_2 = 12$ volts, $R_1 = 10$ ohms, $R_2 = 20$ ohms, and $R_3 = 8$ ohms. Required are the currents in each resistance. Let the currents be as shown in the figure. Since there are but two branch points a and b, there is only one current relation, namely,

$$i_1 = i_2 + i_3 \tag{29}$$

and we need two mesh equations. Consider the mesh $cdabc$. The second Kirchhoff rule yields

$$-E_1 + R_1i_1 + i_3R_3 = 0 \tag{30}$$

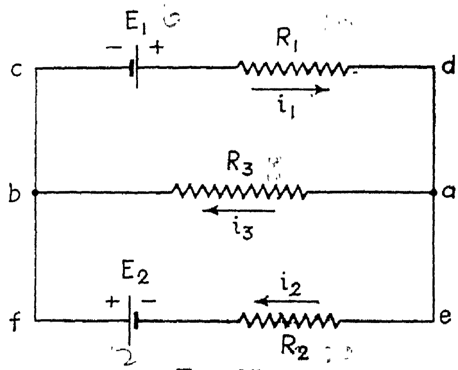

FIG. 35.

and for the mesh $aefba$ (not containing E_1) we obtain

$$i_2R_2 - E_2 - i_3R_3 = 0 \tag{31}$$

Substituting for i_1 in Eq. (30) its value from Eq. (29) there follows:

$$-E_1 + R_1i_2 + (R_1 + R_3)i_3 = 0$$

and inserting numerical values into this equation and Eq. (31) we obtain

$$20i_2 - 8i_3 = 12$$

and

$$10i_2 + 18i_3 = 6$$

from which one obtains readily $i_3 = 0$; $i_2 = 0.6$ amp. and thus $i_1 = 0.6$ amp. In this case no current flows through R_3 and points a and b are at the same potential.

2. The Wheatstone Bridge.—The Wheatstone bridge is a network utilized to measure resistances and is shown in Fig. 36. M and N are fixed resistances, P a variable resistance, and X the unknown. G represents the resistance of a galvanometer and R_B is the internal resistance of the battery of e.m.f. E. Let the currents be as shown. Three independent current relations are

$$i = i_m + i_p \quad \text{(branch point } a) \tag{32}$$
$$i_m = i_n + i_g \quad \text{(branch point } b) \tag{33}$$
$$i_p + i_g = i_x \quad \text{(branch point } d) \tag{34}$$

and three independent mesh equations are

$$Mi_m + Ni_n + iR_B - E = 0 \quad \text{(mesh } abcea) \tag{35}$$
$$Mi_m + Ni_n - Xi_x - Pi_p = 0 \quad \text{(mesh } abcda) \tag{36}$$
$$Mi_m + Gi_g - Pi_p = 0 \quad \text{(mesh } abda) \tag{37}$$

The solution of these six equations yields the current i_g through the galvanometer. The bridge is said to be in balance when the galvanometer current i_g is zero. Under these conditions we have [Eqs. (33) and (34)]

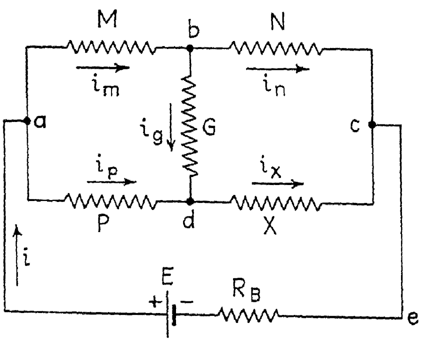

FIG. 36.

$$i_m = i_n$$

and

$$i_p = i_x$$

so that Eqs. (35) to (37) become

$$(M + N)i_m + iR_B - E = 0$$
$$(M + N)i_m - (X + P)i_p = 0$$
$$Mi_m - Pi_p = 0$$

Dividing the second of these equations by the third, one gets

$$\frac{M + N}{M} = \frac{X + P}{P}; \qquad 1 + \frac{N}{M} = \frac{X}{P} + 1$$

or

$$\frac{X}{P} = \frac{N}{M}; \qquad X = P\frac{N}{M} \tag{38}$$

Equation (38) is the condition which must be satisfied to balance a Wheatstone bridge. For a fixed ratio of N to M this may always be accomplished by adjusting P and hence obtaining the value of X.

23. Joule's Law; Power in D.C. Circuits.—Let us consider any d.c. network S, and let a and b represent the terminals of the system across which a definite potential difference V_{ab} is maintained by an external seat of e.m.f. Let the steady current flowing into and out of the system be i (Fig. 37). No detailed knowledge is assumed about the system. Since, by definition, $V_{ab} = V_a - V_b$ is the work done per unit charge (the drop in potential) in moving a charge from a to b and since the charge transported per unit time from a to b is i, it follows that the work done per unit time on S (the power absorbed by the system S) is given by

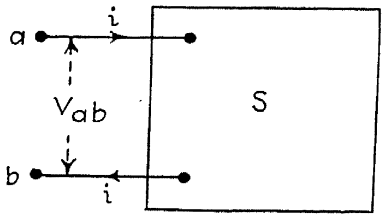

FIG. 37.

$$P = V_{ab} \cdot i \tag{39}$$

This is the general expression for the power input to a d.c. system when a current i is supplied to it and a potential difference V_{ab} is maintained across its terminals. Power in the electrostatic system of units is measured in ergs per second and in the m.k.s. system in joules per second = watts. (1 watt = 1 volt-ampere.)

Just what happens to the energy absorbed by the system S does depend on the make-up of the system. Let us examine some typical cases. Suppose the system S consists solely of a resistance R obeying Ohm's law. In this case Eq. (39) can be extended to

$$P = V_{ab} \cdot i = Ri^2 \tag{40}$$

since by Ohm's law $V_{ab} = Ri$. The term i^2R is *the rate at which heat is evolved* in the resistance, and the statement that the rate of heating of a conductor is equal to i^2R is known as *Joule's law*. As we see from the above, it is totally equivalent to Ohm's law. The equality between power input to a system and the rate of heating is valid *only* if there are no seats of e.m.f. in the system which act as sources of, or as sinks for, energy. To illustrate this, let us suppose that system S consists of a resistance and a storage battery which is being charged (Fig. 38). Let E be the e.m.f.

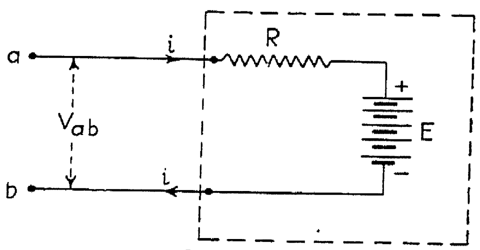

FIG. 38.

of the battery and R the series resistance (including the internal resistance of the battery). For this circuit we have, applying Kirchhoff's rules

$$iR + E - V_{ab} = 0$$

or

$$V_{ab} = E + iR$$

so that the power input to this system is

$$P = V_{ab} \cdot i = Ei + i^2R \tag{41}$$

We see that only a fraction of the input power goes into heat and the remaining term Ei represents the power absorbed by the battery which continually stores up chemical energy (the process of charging). Let us examine this term more closely. The e.m.f. of the battery E was defined as the work done *by* the chemical forces per unit charge in moving a charge from the plate of lower to that of higher potential inside the battery. In the above case in which the current is forced to flow from high to low potential terminal inside the battery, work is done by the external source *against* these chemical forces, so that Ei (the rate of doing work) is the power *absorbed* by the battery. Were the battery discharging, then the expression Ei is the power delivered by a battery to the circuit of which it is a part. To clarify this statement, consider a simple series circuit of a battery of e.m.f. E, internal resistance R_B and external resistance R. For such a circuit we have

$$E = iR + iR_B$$

and multiplying by i,

$$Ei = i^2R + i^2R_B \tag{42}$$

The term i^2R is the rate of heating in the external resistance; i^2R_B is the rate of heating inside the battery, so that Ei represents the total power developed by the battery. To be sure only a fraction (i^2R) is delivered to the circuit *external* to the battery, but that is of no import in this argument.

Returning to Eq. (41), which is often written in the form

$$(V_{ab} - E)i = i^2R$$

The term $-E$ is called the "back electromotive force" in the system S. One often speaks of the net voltage available for

maintaining a current i through the resistance R as the difference between the "applied" voltage V_{ab} and the back e.m.f. E. This terminology is common in discussing motor action in which the rotating armature becomes a seat of "back" e.m.f. In such a case the term Ei represents the mechanical power developed by the motor.

In passing we may point out that we might have logically made Joule's law (rate of heating $= i^2R$) the basis of our discussion of steady currents and derived Ohm's law from it. Resistance would then have been defined by

$$R = \frac{\text{power dissipated}}{i^2} \tag{43}$$

and for direct steady currents the two definitions are identical. In the case of alternating currents, which we shall encounter later, the two definitions cease to coincide. It is then usual to refer to resistance as defined by Eq. (43) as *effective resistance* (R_{eff}) in contrast to the d.c. or "ohmic" resistance.

Problems

1. Two cells of e.m.f. and internal resistance 2 volts, 0.2 ohm and 4 volts, 0.4 ohm are connected in series and the combination is connected to form a simple series circuit with an external resistance of 11.4 ohms.

a. What is the ratio of currents for the two possible connections?

b. What current flows in each case?

2. Sixteen cells, each of e.m.f. E and internal resistance R_B, are connected in a series-parallel arrangement (s cells in series and p of these series combinations in parallel). The whole combination is connected in series with a single external resistance $R = R_B$. Prove that the maximum current which can be sent through the external resistance is four times the current which a single cell would send through it. How many cells are in series for this to happen?

3. Given two batteries, one of e.m.f. 10 volts and internal resistance 0.9 ohm, the other of e.m.f. 3 volts and internal resistance 0.4 ohm.

How must these batteries be connected to give the largest possible current through a resistance R, and what is this current for

a. $R = 0.3$ ohm?

b. $R = 0.4$ ohm?

c. $R = 0.5$ ohm?

4. A storage battery of e.m.f. 24.0 volts and internal resistance 0.5 ohm, is to be charged with a current of 10 amp. The battery in series with a resistance R is connected to a 110-volt power line.

a. Draw a diagram showing the proper connections and mark the polarities.

b. What resistance R is needed?

c. What is the potential difference across the battery terminals during the charging process?

5. How large a resistance must be placed in shunt (in parallel) with a resistance of 1,000 ohms to reduce its resistance by 50 per cent of its original value?

6. A long uniform wire is cut into n equal lengths which are then used to form an n-strand cable. The resistance of the cable is R. What was the resistance of the original wire?

7. A uniform drop wire of total resistance 1,200 ohms is connected to power mains which maintain a potential difference of 120 volts across it. A voltmeter connected between one end of the drop wire and its mid-point reads 50 volts. What is the resistance of the voltmeter?

8. A uniform drop wire of 1,000 ohms resistance is connected across a 100-volt line and a voltmeter of 500 ohms internal resistance is used to measure the potential difference across a portion of the drop wire of resistance R.

Make a plot of the voltmeter reading against R for all possible values of R.

9. A voltmeter is constructed of a galvanometer of 2,000 ohms resistance and two resistances of 2,000 and 6,000 ohms, as shown in Fig. 39. The

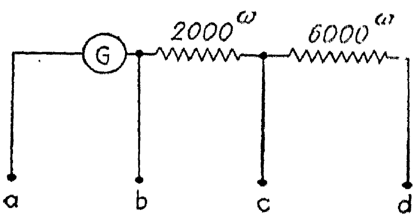

FIG. 39.

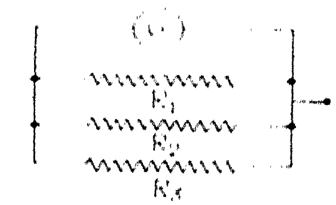

FIG. 40.

galvanometer gives full-scale deflection when a current of 2.5 ma. flows through it. What voltages across the terminals ab, ac, and ad will give full-scale deflections? Would this instrument serve as a suitable ammeter? Explain.

10. An ammeter consists of an 8-ohm galvanometer and three resistances, R_1, R_2, and R_3, all connected in parallel as shown in Fig. 40. By means of a switch R_2, R_3, or R_2 and R_3, may be disconnected. A current of 2 ma. through the galvanometer gives full-scale deflection.

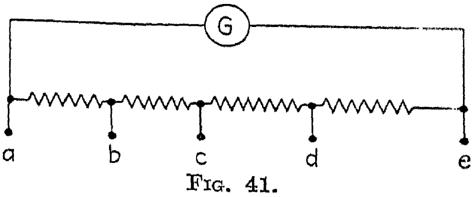

FIG. 41.

What resistances must be used if the ammeter is to have $\frac{1}{10}$-, 1- and 10-amp. scales?

Do the shunts change the percentage accuracy of the instrument?

11. A 10-ohm galvanometer with a 90-ohm resistance permanently connected across its terminals is used as a galvanometer. The instrument may be used with a choice of connections ab, ac, ad, or ae, as shown in Fig. 41. How should the 90-ohm resistance be divided so that the sensi-

tivity (the reciprocal of the current causing full-scale reading) of the instrument should change by a factor of 10 for each successive connection *ab*, *ac*, *ad*, and *ae*?

12. Calculate the resistance of a seven-strand copper cable 2 miles long, each strand being of circular cross section and diameter 0.020 in. The resistivity of copper is 10.4 ohms "per mil-foot."

13. A cylindrical rod of copper of diameter 0.2 in. is drawn into a wire of 10 mils diameter. The resistance of the rod is 0.001 ohm. Calculate the resistance of the wire, assuming that the drawing has no effect on the resistivity of the copper.

14. A cable consists of a central core of steel wire 500 mils in diameter surrounded by a tightly fitting sheath of copper 0.10 in. thick. Calculate the resistance of 1 mile of this cable. What fraction of the current carried in the cable is in the copper? The resistivities of copper and steel on the mil-foot basis are 10.4 and 90 ohms, respectively.

15. How constant must the temperature of a coil of wire be maintained if its resistance is to be constant within 0.1 per cent? The temperature coefficient of resistivity of the metal is 0.004 per degree centigrade.

16. The resistivity of platinum at 0°C. is 54.0 ohms "per mil-foot." Calculate its conductivity in e.s.u. and in the m.k.s. system at 0°C. and at 20°C. The temperature coefficient of resistance of platinum is 0.00354 per degree centigrade.

17. The electrodes of a cell consist of a metal rod 5 cm. in diameter and a coaxial hollow cylinder of inside diameter 25 cm. The electrodes stand on a glass plate in an electrolyte (a conducting solution) which is 20 cm. deep over the glass plate. A current of 5 amp. flows through the cell. Calculate the current density at the surface of each electrode and at any point in the solution.

18. The conducting solution in Prob. 17 has a resistivity of 4.8×10^{-2} ohm-meter (m.k.s.). What is the internal resistance of the cell?

HINT: This problem is analogous to the problem of steady heat flow between coaxial cylindrical surfaces.

19. A current i flows steadily between two concentric spherical electrodes (the spherical condenser), the radius of the inner electrode being r_1 and that of the surrounding spherical surface being r_2. The medium between the spheres has a conductivity σ.

a. Derive an expression for the current density in terms of i at any point between the electrodes.

b. Derive an expression for the resistance between the electrodes.

20. Using Eqs. (32) to (37) of Chap. IV, derive an expression for the galvanometer current i_g when the Wheatstone bridge is *not* balanced.

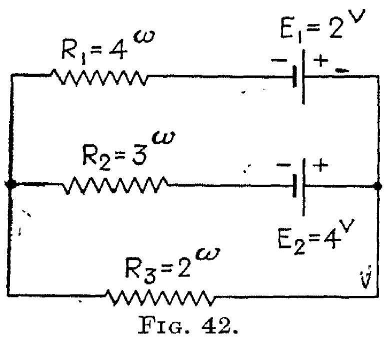

FIG. 42.

21. In the network shown in Fig. 42 find the currents in the three resistances. The battery resistances are included in the values of R_1 and R_2.

22. If the 4-volt battery is reversed in Prob. 21, find the currents in R_2, and R_3.

23. The accompanying diagram [illegible] lustrat [illegible] ire. How [illegible] the slide wire must the con[illegible] be moved in order tha[illegible] va[illegible] the potential difference between the points P and Q when the above adjustment is made?

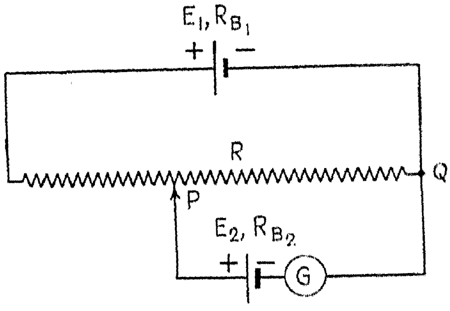

FIG. 43.

24. In the circuit of Fig. 44 find the potential difference between the points a and b. Which is at the higher potential? Battery A has an e.m.f. of 12 volts and an internal resistance of 0.10 ohm and battery B has an e.m.f. of 6.0 volts and an internal resistance of 0.05 ohm.

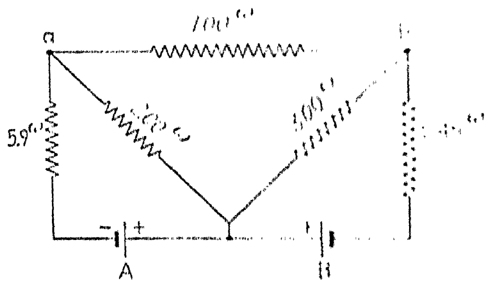

FIG. 44.

25. An electric heater immersed in water raises the temperature of 5,000 grams of water from 0 to 40°C. in 5 min. when a current of 25 amp. flows through the heater. What is the resistance of the heater?

26. A 10-hp. electric motor operates on 110 volts and has an efficiency of 85 per cent. The generator supplying the power has a terminal voltage of 125 volts and is 2,000 ft. from the motor. What diameter of copper wire (ρ = 10.4 ohms "per mil-foot") is needed to deliver current to the motor at its rated voltage?

27. A 72-volt storage battery of 0.33-ohm internal resistance is charged from 110-volt supply mains with a 6-ohm resistance in series with it.

a. How much power is drawn from the mains?

b. What is the rate of heating in the circuit?

c. What is the efficiency of the charging process?

28. A factory draws an amount of power P from a d.c. transmission line at a terminal voltage (at the factory) V. The resistance of the transmission line (both wires) is R.

a. Write an expression for the current drawn and for the generator terminal voltage (in terms of P, V, and R).

b. From the above obtain an expression for the power output of the generating station and also for the percentage of this output delivered to the factory.

CHAPTER V

THE MAGNETIC FIELD OF STEADY CURRENTS

In addition to the heating effects of steady currents which we have studied in the last chapter, there appear mechanical forces of a type quite different from electrostatic forces which are known as "magnetic" forces. It is assumed that the reader is acquainted with the elementary facts concerning the forces which permanent magnets exert on each other. We shall not follow the customary historical treatment of magnetism utilizing permanent magnets as the basis for the discussion because, in spite of the apparent simplicity of the idea of magnetic poles, magnetic phenomena are in fact very different from electrostatic phenomena and the usual analogies drawn between these subjects are apt to lead to serious confusion. We delay a detailed discussion of permanent magnets to a later chapter in which we shall study the magnetic properties of matter and shall introduce the idea of a magnetic field as a field produced by moving charges or electric currents. It is undeniably convenient to visualize a magnetic field existing in a given region of space if a small compass needle freely suspended at its center orients itself in a definite direction (and returns to this orientation if disturbed) and to think of this simple experiment as a practical method of determining the direction of the field at the point where the compass needle is located. Indeed it was as late as 1820 that Oersted discovered that an electric current exerts mechanical forces on a magnet. Immediately thereafter Ampère observed forces of a similar nature between current-carrying conductors and performed a series of fundamental experiments from which the laws of force between currents were derived. Ampère's experiments were performed with currents in conductors, but it was shown later by Rowland that moving charges produce magnetic effects similar to conduction currents and are acted on by forces when in the vicinity of current-carrying circuits or magnets.

24. The Magnetic Induction Vector B.—As stated above, we shall start from the results of the experiments of Ampère, or

their equivalent, to develop the idea of a magnetic field. One of the fundamental facts observed is that the force on an element ds of a conductor carrying a current i due to the presence of a neighboring current-carrying circuit (or a magnet) acts at *right angles* to the length ds. Stated in terms of moving charges, we can say that such a force acts on a moving charge at right angles to the direction of its motion. Any region of space where moving charges (or currents) experience a force of the above type is said to have a magnetic field existing in it. This field is of course due to the presence of other currents or magnets. As in the electrostatic case we shall not concern ourselves at present with the mode of creation of such fields but shall examine the nature of the forces exerted by the field on currents or on moving charges. In order to set up a quantitative description of the field, it will be necessary to introduce the idea of a test body just as we used the idea of a test charge in the discussion of the electrostatic field. As our test body we can employ either an element ds of a conductor carrying a current or, better still, a tiny beam of moving electrons such as one would obtain in a miniature cathode-ray tube. In utilizing such a test body to detect the presence of and to measure a magnetic field, we encounter a complication at the very outset in that the force acting on our current element depends not only on the position of the element but also on its orientation in space, *i.e.*, on the direction of current flow at the point in question. In every case, however, the magnetic force acts at right angles to the current. The magnitude of the force depends among other things on the direction of the current, and there is one direction for which the force becomes zero. *We call this the direction of the magnetic field at the point where the current element is located.* If the current flows at right angles to this direction, the force becomes maximum and for other orientations it is proportional to the sine of the angle between the direction of current flow and the direction of the field. Furthermore, the force is proportional to the current i flowing in the element ds and to the length ds and is always at right angles to the direction of the field as well as to the direction of current flow.

At each point of space we define a magnetic field vector, denoted by B, and called the magnetic induction, *whose magnitude is the maximum force exerted on a current element divided by the*

product of the current and the length of the element and whose direction has been specified on page 82. If we denote this force by dF_{max}, we can write

$$B = \frac{dF_{max}}{i\ ds} \tag{1}$$

as the magnitude of the magnetic induction vector, the direction of this vector being at right angles to both dF_{max} and to ds. As we shall prove in a moment, this is entirely equivalent to

$$B = \frac{F_{max}}{qv} \tag{2}$$

where F_{max} is now the maximum force exerted on a charge q moving with a velocity v, and B is perpendicular to both F_{max} and to v.

If the direction of current flow is not perpendicular to B, then the force on a current element has a magnitude given by

$$dF = i\ ds\ B \sin\theta \tag{3}$$

where θ is the angle between the direction of B and that of ds and dF is normal to the plane determined by B and ds. Similarly, the force F acting on a moving charge is

$$F = qvB \sin\theta \tag{4}$$

where θ is the angle between v and B.

The directions of the vectors whose magnitudes enter into Eqs. (3) and (4) are related in the following manner: If one rotates the vector ds (in the direction of current flow) or v through the smallest possible angle so that it lines up with B, the force points in the direction in which a right-handed screw would move when so rotated (Fig. 45). This rule is very similar to that for finding the direction of the vector moment of a force about a given point, or for finding the direction of the angular momentum of a particle about a given point. The rather cumbersome mode of presentation which has been given can be considerably shortened and clarified

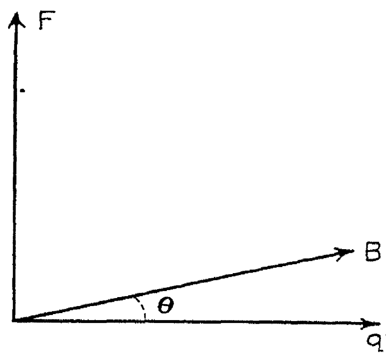

FIG. 45.

by introducing the idea of vector multiplication, and we shall now digress to define these convenient quantities.

Scalar and Vector Products of Vectors. The scalar or "dot" product of two vectors A and B is defined as a scalar of magnitude equal to the product of the magnitudes of A and B and the cosine of the angle between them. It is written in the form

$$\vec{A} \cdot \vec{B} = AB \cos(\vec{A}, \vec{B}) \tag{5}$$

It is clear that the scalar product is equal to the product of the component of one of the vectors in the direction of the other and the magnitude of the second vector. As examples, we have the familiar expression for the work done by a force F in moving a particle through a displacement ds. The work done is

$$dW = \vec{F} \cdot \vec{ds} = F\,ds \cos(\vec{F}, \vec{ds}) = F_s\,ds$$

Similarly the electromotive force around a closed path can be written as a scalar product in the form

$$\cdot \vec{ds} =$$

The second type of product which one forms of two vectors A and B is known as the vector or "cross" product. It is defined as a vector C whose magnitude is the product of the magnitudes of A and B and the sine of the angle between them and whose direction is normal to the plane of A and B. The sense of the vector product of A and B is obtained by rotating A through the smallest angle so that it lines up with B and taking the direction in which a right-handed screw would move as the direction of C. In symbols, one writes

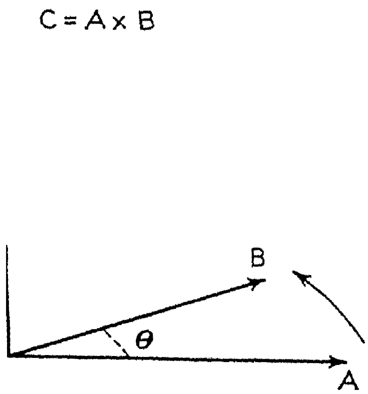

Fig. 46.

$$\vec{C} = \vec{A} \times \vec{B} \qquad \text{(Fig. 46)} \tag{6}$$

and the magnitude of C is $AB \sin\theta$. The order of multiplication is important, and we have

$$\vec{A} \times \vec{B} = -(\vec{B} \times \vec{A}) \tag{6a}$$

In terms of this notation one can write for the torque of a force F about a point O

$$\vec{T} = \vec{r} \times \vec{F}$$

where $\vec{r}$ is the vector drawn from the point O to the point at which the force is applied.

Equations (3) and (4) state that the magnetic force on a current element or on a moving charge can be expressed as the vector product of two vectors. Thus for the current element ds,

$$\vec{dF} = i(\vec{ds} \times \vec{B}) \tag{7}$$

and for the moving charge

$$\vec{F} = q(\vec{v} \times \vec{B}) \tag{8}$$

Equations (7) and (8) are to be looked upon as the equations which define the magnetic induction vector B. We still have to show that they are equivalent.

Consider a straight metallic wire of cross section A carrying a steady current i, and let us suppose that this current is due

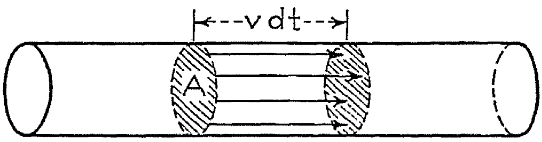

FIG. 47.

to the steady motion of free electrons inside the conductor. Since the electric current is the rate at which electric charge crosses any cross section A of the conductor, it is equal to the product of the charge q on each particle and the number of such particles crossing A per unit time. Now the number of particles crossing A in a time interval dt is equal to the number found inside a cylinder of base A and altitude $v\,dt$, where v is the constant velocity of the charged particles. Thus if n be the number of charged particles per unit volume (a constant for the case of steady flow), the current i may be written as

$$i = nqvA$$

Consider now a length ds of the conductor. We have evidently

$$i\,ds = nqvA\,ds = Nqv$$

where $N = nA\,ds$ is the total number of charged particles in the volume $A\,ds$, so that the magnetic force acting on these particles is, according to Eq. (7)

$$F = i(ds \times B) = Nq(v \times B)$$

and the force acting on *one* particle is this expression divided by N, the number of particles in the element ds. This completes the proof.

We still have to concern ourselves with the question of units. The units of B [according to Eq. (7) or Eq. (1)] depend on the units chosen for force, length, and current. In the m.k.s. system the unit force is 10^5 dynes = 1 newton, the unit length 1 meter, and the unit current is 1 amp. Since the dimensions of force are those of electric field intensity times charge and those of current are charge per unit time, evidently B is expressed in volt-seconds per square meter, sometimes denoted by webers per square meter. B is practically never expressed in e.s.u., so that we need not concern ourselves with that case. There is, however, another system of units called the *electromagnetic absolute system* by virtue of the fact that the units are derived from electromagnetic laws. In this system of units (e.m.u.) the mechanical units are centimeters, grams, and seconds and the unit current, called the *abampere*, is equal to 10 *amp*. The size of the unit of B in e.m.u. is determined by the defining equation (1) or (7), and it is called 1 *gauss*. In the electromagnetic system the unit charge is called 1 abcoulomb and equals 10 coulombs, the unit potential 1 abvolt = 10^{-8} volt, etc. The units in the e.m.u. are related to those of the m.k.s. system by powers of ten. As we shall see later when we define the abampere precisely, this is a consequence of definition and not of experiment.

25. Magnetic Flux; Solenoidal Nature of the Vector Field of B. At each point of space we can imagine the vector B constructed; the totality of these vectors in a given region is called the magnetic field there. One can construct lines of B which give the direction of B at every point and can limit the number of lines per unit area crossing a tiny area which is normal to the field

in such a way that the number of lines per unit area is made equal to the numerical magnitude of B. This is the same convention as was employed for the electrostatic field. The total number of lines of magnetic induction crossing an area is called the *magnetic flux* across the area, and we have evidently

$$\Phi = \int B_n \, dS \tag{9}$$

Φ is the magnetic flux, and B_n is the component of B normal to the surface at the point where dS is located (Fig. 48). From Eq. (9) we see that the magnetic induction B can be measured

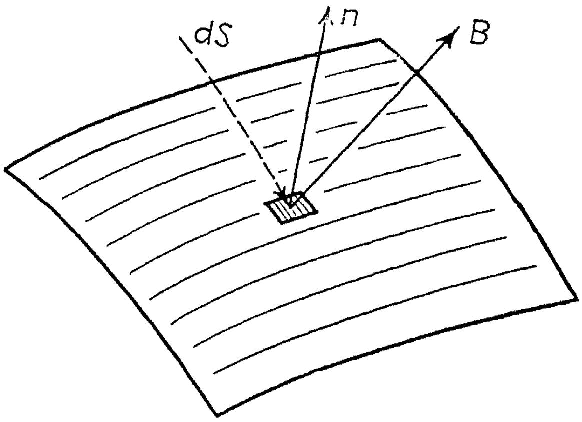

FIG. 48.

in flux units per unit area, and hence it is often called *magnetic flux density*. The unit of flux in the electromagnetic system of units is called 1 *maxwell*, so that 1 gauss = 1 maxwell/cm.^{2} In the m.k.s. system the unit of flux is called 1 *weber* so that the unit of B is 1 weber/meter^2 as stated in the previous section.

The nature of the vector field of B is *radically different* from that of the electrostatic field. In the electrostatic case we saw that the lines of D or $\mathcal{E}$ (in vacuum) always terminate on charges and hence could never form closed curves. Furthermore, one could introduce a scalar potential having a definite value at each point of space from which the field could be obtained by differentiation. These facts are *not* true for the field of B. *The lines of B can never start or stop at any point and hence always form closed curves.* A vector field of this sort is called solenoidal or source-free, since no starting points or sources can be assigned to the lines describing the field. We have already encountered one such field in our study of steady currents, since the conserva-

tion of charge requires that the lines of electric current flow be closed in the steady state (the equation of continuity). Thus one can say that the vector field describing the steady flow of electric currents (the field of j) is solenoidal. If we wish to express the above facts for B mathematically, we note that the *total* magnetic flux crossing any *closed* surface must be zero since no lines can start or stop inside the volume enclosed by the surface. In symbols, this becomes

$$\int_{\text{surface}} B_n \, dS = 0 \tag{10}$$

Equation (10) *is one of the fundamental laws of electromagnetic theory.* Furthermore, one cannot introduce a scalar potential to describe such a field. It can be shown that it is impossible to find a scalar function of position such that a purely solenoidal vector field may be obtained from it by differentiation, *i.e.*, by taking the gradient of a scalar function.

23. Motion of Charged Particles in Magnetic Fields.—In this section we shall investigate the motion of charged particles both in magnetic and in combined electric and magnetic fields. In so doing we must remember that we assume that we can neglect the modification of the field caused by the presence of the moving charges. This condition is often realized in practice when the magnetic field produced by these moving charges is negligible compared to the external field in which they move. The study of the motion of ions or of electrons, in particular the determination of the orbits, is of fundamental importance since the methods of determining electronic and ionic masses are based on such a study.

From the fundamental Eq. (8) giving the force on a moving charge as $F = q(v \times B)$, it is clear that, since the force F is normal to the direction of motion, the kinetic energy of the particle is not affected by the magnetic field and only the direction of v changes. For this reason the magnetic force is often called a *deflecting* force. Let us consider the problem of the motion of a particle which is projected with an initial velocity v_0 into a region of space where there exists a uniform field B, the direction of v_0 being at right angles to B. The subsequent motion of the particle is one of constant speed v_0 in a curved path,

the component of force along the path being zero. Normal to the path, in a direction n, let us say, we have from Newton's second law

$$F_n = \qquad = ma_n =$$

since the angle between the direction of the motion and B is 90° and the component of acceleration normal to the path is v_0^2/r, with r the radius of curvature of the path. Solving for r, we obtain

$$r = \frac{mv_0}{qB} \tag{11}$$

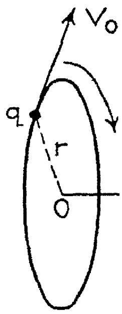

FIG. 49.

and as all the quantities on the right-hand side are constant, the path is a circle of radius r, the circle being a curve of constant radius of curvature (Fig. 49). The direction of rotation is as shown for a positively charged particle, the opposite direction for a negative charge. The angular velocity ω is

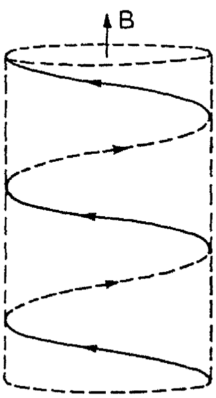

FIG. 50.

$$m$$

and hence the period T is

$$T = \frac{2\pi}{\omega} = \qquad q \qquad B \tag{12}$$

The time for one complete rotation does not depend on the velocity of particle! This is a very important fact and much use is made of it in the experimental methods of atomic physics.

If the particle has an initial velocity v_0 not perpendicular to B, we can resolve v_0 into two components, one perpendicular to and one parallel to B. The latter component is not affected by the magnetic induction; hence the motion is a superposition of circular motion described by Eq. (11) in a plane normal to B and a uniform translation parallel to B. The path is a helix of uniform pitch (Fig. 50). One important use of these results is in the application to magnetic focusing by a longitudinal magnetic field. In Fig. 51 is shown a slit S through which ions enter a region where there is a uniform magnetic field B. All the ions

entering at an angle β with the induction B will be brought to a focus at P on a fluorescent screen S' located at a distance d from the slit if the time of flight from S to P is just equal to the time for one revolution in the helical path. The time of flight is

$$t = \frac{d}{v_0 \cos \beta}$$

and, equating this to the expression given by Eq. (12), there follows

$$\frac{d}{v_0 \cos \beta} = \frac{2\pi m}{qB}$$

The focal length d depends on the initial velocity v_0, the angle β,

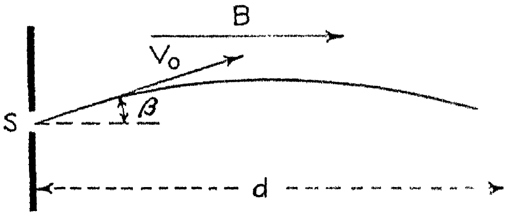

FIG. 51.

the induction B, and the ratio of charge to mass of the particles.

The ratio of charge to mass of an ion may be determined for a known initial velocity of the ion by deflecting the ions by a con-

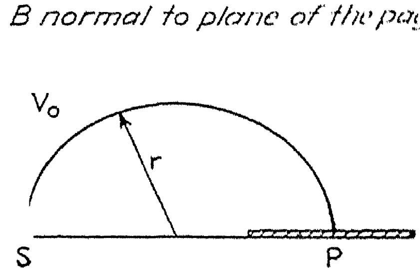

FIG. 52.

stant magnetic field normal to the direction of motion, as shown in Fig. 52. The ions entering through the slit S are deflected in a semicircle and impinge on a photographic plate at P. The distance SP is given by

$$SP = \frac{\quad}{qB}$$

and v_0 can be determined by allowing the ions to fall through a known electrostatic potential drop V before entering the slit S. Thus the ratio q/m can be determined and this is the principle underlying the operation of a mass spectrograph.

In making numerical calculations one must be careful to express charge in abcoulombs (e.m.u.) when B is expressed in gauss, or in coulombs when B is expressed in webers per square meter (m.k.s.). In calculations concerning the motion of charged particles in combined electric and magnetic fields, particular care must be employed since the electric field intensity is so often expressed in e.s.u. The general expression for the force on a charge q moving in a combined electric and magnetic field is

$$F' = q\mathcal{E}' + q(v' \times B') \tag{13}$$

This equation can be used as it stands if *all* quantities entering into it are expressed in the same system of units, *e.g.*, in the m.k.s. system. The custom of expressing electric intensity and charge in e.s.u. and magnetic induction in gauss (e.m.u.) in the same equation is widespread (the mechanical units are those of the c.g.s. system in both e.s.u. and e.m.u.), and, if this is done, Eq. (13) *cannot* be used as it stands. This mixed system of units, partly electrostatic and partly electromagnetic, is called the *Gaussian* system. To find the correct form of Eq. (13), we can write

$$\begin{array}{ccccccccc} F & = & q & \mathcal{E} & + & q & (v & \times & B) \\ \downarrow & & \downarrow & \downarrow & & \downarrow & \downarrow & & \downarrow \\ \text{(dynes)} & & \text{(e.s.u.)} & \text{(e.s.u.)} & & \text{(e.m.u.)} & \text{(cm./sec.)} & & \text{(gauss)} \end{array}$$

and if we denote the ratio $\dfrac{q \text{ (in e.s.u.)}}{q \text{ (in e.m.u.)}}$ by c, the above equation becomes

$$F \text{ (dynes)} = q \text{ (e.s.u.)}\left\{\mathcal{E} \text{ (e.s.u.)} + \frac{1}{c}[v \text{ (cm./sec.)} \times B \text{ (gauss)}]\right\} \tag{14}$$

The numerical value of c as determined experimentally is very nearly 3×10^{10}. We shall return to a discussion of its dimensions later.

27. Side Thrusts on Conductors; the Moving-coil Galvanometer.—The magnetic forces exerted on current-carrying wires are generally termed "side thrusts" because they act normally to the length of the conductor. We have formulated the law for this force in the differential form

$$d\vec{F} = i(d\vec{s} \times \vec{B})$$

and, in order to obtain the resultant force (or torque) acting on any finite length of a conductor, the above equation must be integrated. Since it is a vector equation, this must be done in general by computing the components of dF before integrating.

Steady currents always flow in closed conducting paths; hence it is important to consider the magnetic force (and torque) on a closed loop of wire carrying current. We shall employ a loop in the form of a rectangle for simplicity, but the results will be stated in a manner which will be valid for an arbitrary shape of loop. In a uniform field of magnetic induction B, a closed current loop experiences *no* magnetic force. Only if the field is inhomogeneous is there a resultant force on the loop. The forces acting on the sides of a rectangular loop whose plane is normal to a uniform field B are shown in Fig. 53. The upward force on the side ab is just balanced by the downward force on side cd. Similar balance occurs between the forces on the vertical sides of the loop ad and bc, as shown. Evidently this cancellation of forces occurs even if B is not normal to the plane of the loop. In this case the magnitude of the side thrust on ab would equal $iB \sin \alpha(ab)$, where α is the angle between B and the current direction in the length ab. The side thrust on cd is similarly $iB \sin \alpha(cd)$ but of opposite sign since the current flows in a direction opposite to that in ab. Since $cd = ab$, we again have a balance. In a nonuniform field of B, however, this cancellation does not occur, as B or $\sin \alpha$, or both, will vary from point to point on the loop so that in general there will be a resultant force.

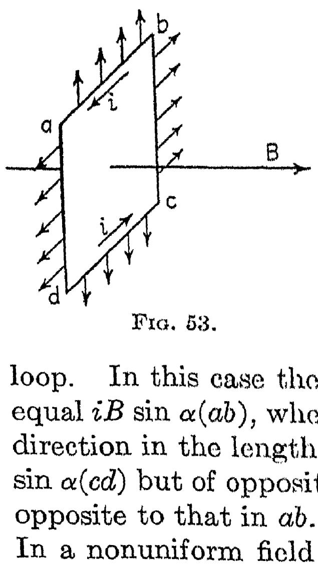

FIG. 53.

We next consider the case of a rectangular loop of wire carrying a current i which is free to rotate about a horizontal axis AA passing through its center, as shown in Fig. 54a. The width of the loop is w, its length is l, and let the current i flowing around it be as shown. Suppose this loop is in a uniform magnetic field normal to the axis of rotation. The two equal and opposite forces F act on the two conductors of length w vertically. These forces give rise to a torque about the axis AA. The forces

acting on the other two sides of the loop give rise to no torque about this axis. In Fig. 54*b* is shown the same loop, as one looks at it along the axis AA. We see that the torque about the axis is

$$F\frac{l}{2}\sin\theta + F\frac{l}{2}\sin\theta = Fl\sin\theta \quad .$$

where θ is the angle between the normal n to the plane of the loop and B. Since F is the force acting on a conductor of length w

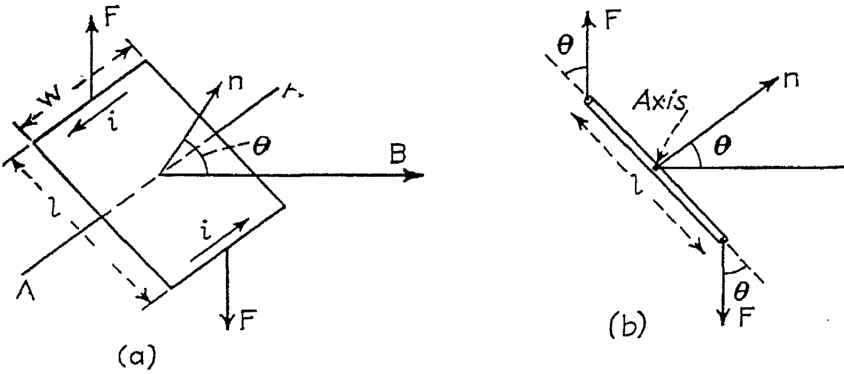

FIG. 54.

and B is everywhere constant and at right angles to the current flowing in this conductor, the force F is

$$F = Bwi \tag{15}$$

so that the torque becomes

$$T = Bi(wl)\sin\theta$$

and, since wl is the area of the loop, this can be written as

$$T = BiA\sin\theta \tag{16}$$

This torque is a torque tending to rotate the loop into a position in which its plane is normal to B. In this position $\theta = 0$, and the torque T is also zero. It is a position of stable equilibrium. Equation (16) turns out to be true for a loop of arbitrary shape, the torque being proportional to the area of the loop. The maximum torque occurs when $\theta = 90°$, *i.e.*, when the normal to the plane of the loop is perpendicular to B.

One of the most common types of galvanometers, the so-called d'Arsonval or moving-coil galvanometer, functions by virtue of the torque action discussed above. A coil of N turns in series is suspended by wires W, as shown in Fig. 55. For the uniform

field B shown in the figure, one of the coil sides is pushed toward and the other away from the reader, giving rise to a torque about the axis of suspension

$$T = NBiA \tag{17}$$

The magnetic field is supplied by a permanent magnet, which in actual design is so arranged that B is in the plane of the coil even when the latter is twisted through an angle (a so-called radial field). When current flows through the coil, it rotates through an angle θ until the torsional restoring torque of the suspension is equal and opposite to the magnetic torque. Since the restoring torque is proportional to θ, we can write for equilibrium

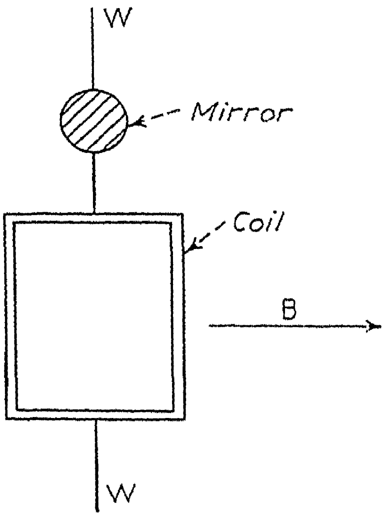

FIG. 55.

$$NBiA = k'\theta \tag{18}$$

where k' depends on the torsion modulus, length, and radius of the suspending wires.

Equation (18) may be written in the form

$$\frac{NBA}{k'}\theta = \text{constant} \times i \tag{19}$$

so that the angular deflection is proportional to the current. The discussion of the sensitivity of the instrument is left to the problems.

The principle underlying the action of an electric motor is essentially that for the moving-coil galvanometer, the only essential difference being that the coil (more precisely, the armature carrying the coil) is free to rotate about its axis in bearings instead of being subjected to a restoring torque.

28. Ampère's Rule; The Magnetic Intensity H.—We now must inquire into the question of the laws governing the production of magnetic fields by currents. These laws were discovered, as previously mentioned, by Ampère as a result of his experiments on the forces exerted by one current-carrying conductor on another. Just as in the electrostatic case, one finds that these magnetic forces depend not only on the currents but also on the material bodies which are present. Once again we defer the treatment of the effects of the boundaries between such material

media and shall confine ourselves in this chapter to the case of *infinite homogeneous media* (principally empty space); hence the boundary effects may be entirely neglected. The result of Ampère's experiments may be stated as follows: The magnetic induction vector B at a given point of space may be considered as the vector sum of infinitesimal vectors dB, each of the latter being due to a current element $i\,ds$ of the circuit producing the field. Let $\vec{r}$ be the radius vector from the current element $i\,\vec{ds}$ at O to the point P where the field B is being calculated (Fig. 56). The magnitude of dB due to this current element is proportional to the magnitude of $i\,ds$, to the sine of the angle α between $i\,ds$ and r, and inversely to the square of the distance r. In symbols,

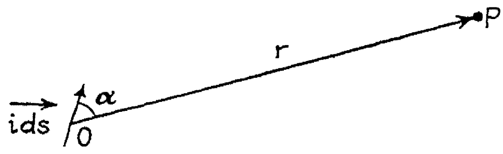

FIG. 56.

$$dB \cdot \quad i\,ds \sin\alpha$$

The direction of dB is at right angles both to ds and to r, and, if the vector $\vec{ds}$ is rotated through the smallest possible angle so that it lines up with $\vec{r}$, the direction in which a right-handed screw would move is the direction of $\vec{dB}$. Using the notation for a vector product, we may write the above proportionality in the form

$$\vec{dB} \sim \frac{i(\vec{ds} \times \vec{r})}{r^3}$$

If there is a current element $i_1\,ds_1$ at P, the force on it is, according to Eq. (7),

$$\vec{dF} = i_1(\vec{ds}_1 \times \vec{dB}) \sim \frac{i_1 i}{r^3}\,\vec{ds}_1 \times (\vec{ds} \times \vec{r}) \qquad (20)$$

The proportionality expressed by the above relation summarizes Ampère's law for the force exerted on a current element $i_1\,ds_1$ by another current element $i\,ds$ separated from it by a distance r. If we denote the proportionality factor by μ, the above proportionality can be written as an equation

$$\overrightarrow{dF} = \mu \frac{i i_1}{r^3} \overrightarrow{ds_1} \times (\overrightarrow{ds} \times \vec{r}) \tag{21}$$

The proportionality factor μ is called the *magnetic permeability* of the medium in which the conductors are embedded. It is the only factor in Eq. (21) depending on the medium, the other factors depending only on the currents and their relative positions. Remember that this statement is valid *only* for infinite homogeneous media or in such cases for which boundary effects are negligible. Thus we can write for $\overrightarrow{dB}$, due to the current element $i\,\overrightarrow{ds}$ (Ampère's rule),

$$\overrightarrow{dB} = \mu i \frac{\overrightarrow{ds} \times \vec{r}}{r^3} \tag{22}$$

and, in the case for which the medium is empty space (or, for all practical purposes, air), we write

$$\overrightarrow{dB} = \mu_0 i \frac{\overrightarrow{ds} \times \vec{r}}{r^3} \tag{23}$$

where μ_0 is the magnetic permeability of empty space.

Just as in the case of electrostatics, it is convenient to introduce a second magnetic vector which depends on the currents producing the field and which shall be independent of the medium in which the field is produced. This vector (the analogue of D in electrostatics) is called the *magnetic intensity* and is denoted by the letter H. The contribution to H at a given point of space due to a current element $i\,\overrightarrow{ds}$, is *by definition* for an infinite medium

$$\overrightarrow{dH} = i \frac{\overrightarrow{ds} \times \vec{r}}{r^3} \tag{24}$$

and the vector $\vec{H}$ is obtained by summing the elementary vectors $\overrightarrow{dH}$ produced by all the current elements. The magnitude of the elementary vector $\overrightarrow{dH}$ is accordingly

$$dH = \frac{i\,ds \sin \alpha}{r^2} \tag{25}$$

Comparing Eq. (24) with Eq. (22) and with Eq. (23), we see that

$$H = \frac{1}{\mu} B \tag{26}$$

or, for vacuum,

$$H = \frac{1}{\mu_0} B \tag{27}$$

Thus far we have not considered the question of units. There are two sets of units commonly employed in magnetic calculations. First, the *electromagnetic absolute system of units* (e.m.u.) is characterized by the fact that the permeability μ is arbitrarily chosen as a pure member without dimensions and the value of μ_0 is taken as unity. (Compare this mode of procedure with that for defining ϵ in electrostatics.) Furthermore, c.g.s. units are used for all mechanical quantities just as in the electrostatic system. Since μ is dimensionless, we see from Eq. (21) that the dimensions of current and hence of charge can be expressed in terms of mass, length, and time. Thus this equation shows that current in e.m.u. has the dimensions given by

$$[i^2]_{\text{e.m.u.}} = mlt^{-2}$$

and hence charge q in e.m.u. has dimensions given by

$$[q^2]_{\text{e.m.u.}} = ml$$

In sharp contrast to this we have for the dimensions of q^2 in e.s.u., according to Coulomb's law,

$$[q^2]_{\text{e.s.u.}} = ml^3t^{-2}$$

from which there follows the surprising fact that the dimensions of charge in e.s.u. and in e.m.u. are different from each other. This has been a source of great confusion to the student of electromagnetic theory. In fact the ratio of the dimensions of q in e.s.u. to those of q in e.m.u. is readily seen to be

$$\frac{[q_{\text{e.s.u.}}]}{[q_{\text{e.m.u.}}]} = \frac{l}{t} \qquad \text{(a velocity!)} \tag{28}$$

and is expressed in centimeters per second. Equation (21) applied to two current elements in vacuum defines the unit current in e.m.u., the so-called abampere. For simplicity, let us suppose that the two current elements are parallel to each other carrying the same current. Expressing all quantities in e.m.u., Eq. (21) becomes

$$dF = \mu_0 \frac{i^2\, ds_1\, ds}{r^2} \tag{29}$$

and is a force of attraction acting along r (Fig. 57). The element $i\,ds_1$ produces a field at P normal to and into the paper,

$$dB = \frac{\mu_0 i\,ds_1}{r^2} \qquad (\mu_0 = 1)$$

and hence the force on $i\,ds$ is as shown and of magnitude

$$dF = i\,ds\,B = \mu_0 \frac{i^2\,ds\,ds_1}{r^2} \qquad (\mu_0 = 1)$$

Thus, if two parallel current elements of equal length ds attract each other with a force of 1 dyne per unit length (centimeter) of each element when separated by a distance of 1 cm., the current i in each is 1 *abampere.* This is not a convenient working definition, although logically correct, since one always has closed circuits; hence we shall give an equivalent, more convenient definition later.

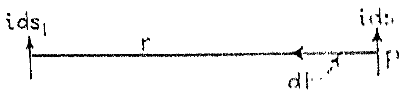

FIG. 57.

The ratio of a charge (or a current) expressed in e.s.u. to the same charge (or current) expressed in e.m.u. turns out to be experimentally

$$\frac{q\text{ (in e.s.u.)}}{q\text{ (in e.m.u.)}} = \frac{i\text{ (in e.s.u.)}}{i\text{ (in e.m.u.)}} = c = 3 \times 10^{10}\text{ cm./sec.}$$

According to Eq. (27) the vectors H and B become identical for vacuum when e.m.u. are employed. In isotropic material media they differ in magnitude but have the same direction and dimensions.

In the m.k.s. system of units, the dimensions of μ or μ_0 are fixed by Eq. (21) in terms of mass, length, time, and charge. Thus we have

$$[\mu] = \frac{ml}{q^2}$$

The unit charge in the m.k.s. system is *defined* as one-tenth the e.m.u. unit and is called the coulomb. Thus we have 1 coulomb = $\frac{1}{10}$ abcoulomb and 1 amp. = $\frac{1}{10}$ abampere. The magnitude of μ_0 in m.k.s. units can be readily calculated from Eq. (29), and is

$$\mu_0 = 10^{-7}\text{ kg.-meter/coulomb}^2 \tag{30}$$

Let us now summarize the discussion of this section. Suppose we have the problem of determining the force exerted on one current-carrying circuit by another. The following procedure is to be followed:

1. Calculate the vector H, due to the second circuit at a point where an element of the first circuit is located, by applying Eq. (24) and by integrating over all the current elements of the second circuit. Since H is a vector, we must first take components of dH, integrate separately for each component, and then find the resultant H.
2. Find the vector B from $B = \mu H$.
3. Calculate the force on the element of the first circuit due to this induction vector B in accordance with Eq. (7).
4. Integrate over all elements of this first circuit to find the resultant force. This calculation is to be carried out by first taking components of the force and then integrating as in procedure 1.

In the following sections we shall concern ourselves mainly with part one of the above scheme and confine our attention exclusively to the consideration of magnetic fields in vacuum or in air, so that $B = \mu_0 H$.

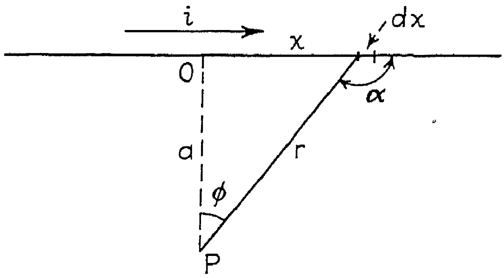

FIG. 58.

29. The Biot-Savart Law; Examples.—As a first example of the application of the laws expounded in Sec. 28, let us consider the magnetic field of a very long straight wire carrying a steady current i (Fig. 58). From the symmetry of the problem we see that the field intensity H can depend only on the distance a of the field point P from the wire. The contribution to H from any current element $i\,dx$ is a vector pointing straight *into* the page at P, at right angles to the element $i\,dx$ and to r. Since all the vectors dH have the same direction (only one component of H), we may integrate directly and have from Eq. (24)

$$H = \quad dx \sin \alpha \tag{31}$$

It is simplest to use the angle ϕ as an integration variable. From the figure it is evident that $\sin\,\alpha = \sin\,(\pi - \alpha) = \cos\,\phi$;

$x = a \tan \phi$; $dx = a \sec^2 \phi \, d\phi$; $\frac{1}{r} = \frac{\cos \phi}{a}$; $\frac{dx}{r^2} = \frac{d\phi}{a}$, so that Eq. (31) becomes

$$H = \frac{i}{a}\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \cos \phi \, d\phi = \frac{i}{a}\Big| \sin \phi \Big|_{-\frac{\pi}{2}}^{\frac{\pi}{2}} = \frac{2i}{a}$$

Thus we have for the magnetic intensity H due to a long straight wire at a point P a vector of magnitude

$$H = \frac{2i}{a} \tag{32}$$

which is directed at right angles to the plane of the wire and a. This result is known as the Biot-Savart law and is of great

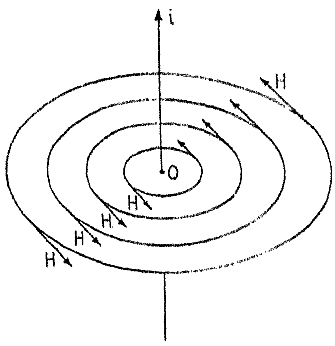

FIG. 59.

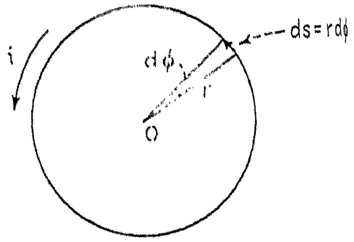

FIG. 60.

importance, since it gives the field near a straight wire of *finite* length if the distance a is small compared with the length of the wire.

Let us imagine that the point P of Fig. 58 describes a circle of radius a about the wire with the center at O. At every point of this circle, the vector H (and also B) is tangent to it and hence the circle is one of the lines of H or B. Thus we see that the lines of magnetic intensity or induction produced by a long straight wire are circles with their centers at the wire (Fig. 59). Note that, if one moves around a line of magnetic intensity in the direction of H, the current producing the field is in the direction in which a right-handed screw would move when so rotated.

Let us now examine the magnetic field produced by a circular loop of wire carrying a steady current i. First, we calculate the field H at the center of the loop (Fig. 60). The field intensity

H is normal to the plane of the loop at O and points out at the reader. For all elements ds, the angle between ds and r is 90°, so that Eq. (25) becomes simply

$$dH = \frac{i\,ds}{r^2}$$

and, since r is constant, we have

$$H = \frac{i}{r^2}\int ds = \frac{2\pi i}{r} \tag{33}$$

We can immediately extend this law to calculate the magnetic intensity at any point on the axis of the loop (Fig. 61). The

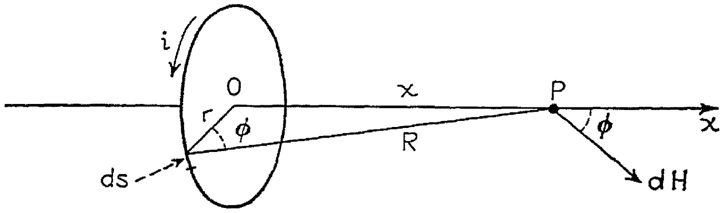

FIG. 61.

distance R from any element ds of the loop to P is constant and is given by

$$R = \sqrt{x^2 + r^2}$$

Furthermore, ds is normal to R at all points of the loop so that in Eq. (25) $\sin \alpha = 1$. Hence

$$dH = i\frac{ds}{(x^2 + r^2)}$$

Before integrating we must resolve H into a component along the x-axis and one at right angles to it. By symmetry the latter vanishes for the whole circuit and $H = H_x$. Thus we have

$$dH_x = dH \cos \phi = dH \cdot \frac{r}{R}$$

and

$$H_x = i\int \frac{r\,ds}{(x^2 + r^2)^{\frac{3}{2}}} = \frac{ir}{(x^2 + r^2)^{\frac{3}{2}}}\int ds = \frac{2\pi i r^2}{(x^2 + r^2)^{\frac{3}{2}}} \tag{34}$$

This expression for the axial field of a circular loop of current is useful for calculating the axial field of any number of coaxial circular loops. As an example, we consider the case of a *solenoid*,

which is a closely wound helical coil of conducting wire. We can get the axial field by integrating the contributions from the individual turns as given by Eq. (34), since for very close winding each turn is almost exactly circular.

We replace the solenoid by a current flowing circumferentially around a hollow cylinder (everywhere perpendicular to the axis). If there are n turns per unit length of the solenoid, each carrying

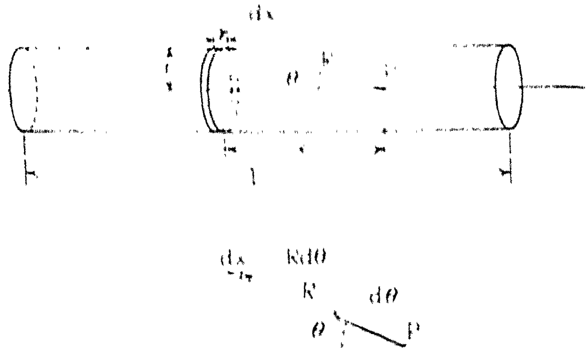

FIG. 62.

a current i, then the current per unit length is ni and the current carried in a circular ring of thickness dx is $ni\,dx$ (Fig. 62). The contribution of this ring to the field intensity at P is according to Eq. (34)

$$dH = \frac{2\pi r^2 ni\,dx}{(x^2 + r^2)^{\frac{3}{2}}}$$

where x is the distance from the plane of the circle to P. This expression must be integrated over the whole length l of the solenoid to find the axial field H at P. From the lower sketch in Fig. 62 we see that

$$R\,d\theta = \sin\theta\,dx$$

and hence

$$dx = R\,\frac{d\theta}{\sin\theta}$$

and since

$$\frac{r}{R} = \frac{r}{(x^2 + r^2)^{\frac{1}{2}}} = \sin\theta,$$

our expression for dH becomes

$$dH = \frac{2\pi nir^2\,d\theta \cdot R}{(x^2 + r^2)^{\frac{3}{2}}\sin\theta} = 2\pi ni\sin\theta\,d\theta$$

The integration immediately yields

$$H = 2\pi ni \int_{\theta_1}^{\theta_2} \sin\theta\, d\theta = 2\pi ni\,(\cos\theta_1 - \cos\theta_2) \tag{35}$$

where θ_1 and θ_2 are the values of θ when dx is at the ends of the solenoid.

When the solenoid is of infinite length, $\theta_1 = 0$ and $\theta_2 = \pi$, so that

$$H = 4\pi ni \tag{36}$$

is the magnetic intensity on the axis of an infinitely long solenoid wound with n turns per unit length, each turn carrying a current i. Equation (35) can be used to find the axial field inside a solenoid of finite length l. Further discussion is left to the problems.

30. The Ampère Circuital Law for H.—There is a more general relation, also due to Ampère, than Eq. (24) between the magnetic

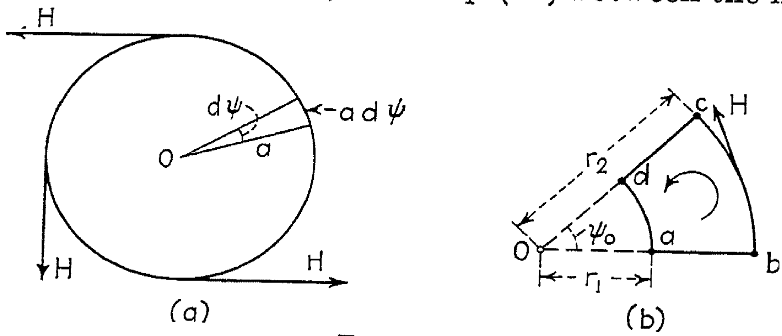

FIG. 63.

intensity H and the steady current i which produces it, and this relation is known as the *circuital law*. To formulate this, we must introduce the idea of the magnetomotive force (m.m.f.) along a path. This is defined as $\int H_s\, ds$ taken along the given path. If the path is closed, we write it as $\oint H_s\, ds$, the circle indicating the fact that a closed path is employed. *The integral is always to be evaluated by traversing the path in such a direction that the enclosed area is kept on the left.*

Let us consider the expression for the magnetomotive force around a closed path in the field of a long straight wire, given by the Biot-Savart law, Eq. (32). First we take the closed path as one of the closed lines of H (a circular path) as shown in Fig. 63*a*. Since H is everywhere tangential to the circle, we have

$$\text{m.m.f.} = \oint H_s\, ds = \oint H\, ds = \oint Ha\, d\psi$$

If now from Eq. (32) we insert the value $H = 2i/a$ into the preceding integral, there follows

$$\oint H_s\,ds = \oint \frac{2i}{a} \cdot a\,d\psi = 2i\oint d\psi = 4\pi i \tag{37}$$

since the integral of $d\psi$ around the circle is just 2π. It can be shown in general that the $\int H_s\,ds$ around *any* closed path which encloses the wire has the value $4\pi i$. On the other hand, if the closed path does not enclose the wire, the integral is zero. We can readily check this last statement for a closed path, the path consisting of two radial sections and two arcs of circles of radii r_1 and r_2, as shown in Fig. 63*b*. We have

$$\oint H_s\,ds = \int_a^b H_s\,ds + \int_b^c H_s\,ds + \int_c^d H_s\,ds + \int_d^a H_s\,ds$$

The first and third terms are zero since H is everywhere normal to a radius vector. The second integral is

$$\int_b^c H_s\,ds = \int_{\psi=0}^{\psi=\psi_0} \frac{2i}{r_2} r_2\,d\psi = 2i\psi_0$$

and the fourth one is

$$\int_d^a H_s\,ds = \int_{\psi_0}^{0} \frac{2i}{r_1} r_1\,d\psi = -2i\psi_0$$

Thus we have shown that $\oint H_s\,ds = 0$ for this path which encloses no current.

The results illustrated above turn out to be true in general and hold not only for steady current flow in wires but also for currents distributed in space. This is the essence of *Ampère's circuital law*, and it is one of the fundamental laws of electromagnetic theory. The formal statement of this law is as follows:

The magnetomotive force around any closed path is equal to 4π times the current crossing any surface of which the closed path is a boundary.

This law has unrestricted validity, holding in all cases, whereas the Ampère rule as expressed in Eq. (24) is in general valid only for media of infinite extent. For the purposes of the present chapter, however, the Ampère rule is more convenient, since the complete determination of the magnetic field requires the use of both the Ampère circuital law and Eq. (10). The more funda-

mental character of the circuital law will become evident when we study the magnetic behavior of matter and investigate the effects of boundaries between magnetic media.

In symbols we write the circuital law in the form

$$\oint H_s\,ds = 4\pi i \tag{38}$$

If i is expressed in abamperes, then H is also in e.m.u. (abamperes per centimeter), whereas, if i is expressed in amperes, ds must be expressed in meters to obtain H in m.k.s. units (amperes per meter). If one wishes to use the Gaussian mixed system expressing H in e.m.u., ds in centimeters, and i in statamperes (e.s.u.), the above equation must be written in the form

$$\oint H_s\,ds = 4\pi\frac{i}{c} \tag{38a}$$

One must not lose sight of the fact that this law as written holds *only* for steady currents, since in this case the lines of current flow are closed and hence the current crossing any surfaces having a common perimeter is the same for all such surfaces.

If the current flow is distributed throughout space, we have for the current flowing across any surface

$$i = \int j_n\,dS$$

so that the Ampère circuital law takes the general form

$$\oint H_s\,ds = 4\pi\int j_n\,dS \tag{39}$$

where the surface integral is taken over *any surface having the closed path as a boundary.*

As an example of an application of Eq. (38) let us consider the magnetic field *inside* a long straight wire of radius R carrying a steady current i. The lines of H inside the wire will be circles concentric with those outside the wire, but H will not vary as $1/r$, r being the distance from the central axis of the wire. To find the correct variation, we evaluate the magnetomotive force around a circle of radius r less than R and have (by symmetry H can vary only with r and hence is constant along the circle and tangential to it at every point)

$$\oint H_s\,ds = 2\pi r H$$

This must be 4π times the current flowing across the area enclosed by the circle. If the current density is $j = i/\pi R^2$, this current is simply $j\pi r^2$, so that

$$2\pi r H = 4\pi(j\pi r^2) = 4\pi i \frac{r^2}{R^2}$$

and solving for H, there follows

$$H = \frac{2ir}{R^2} \qquad (r < R) \tag{40}$$

For $r > R$, we still have $H = 2i/r$, and we note that there is no discontinuity in H as one passes out of the wire since both Eq. (40) and Eq. (32) yield the same value $2i/R$ at the surface of the wire.

The calculation of the magnetic intensity inside a *very long* closely wound solenoid can be accomplished readily with the help

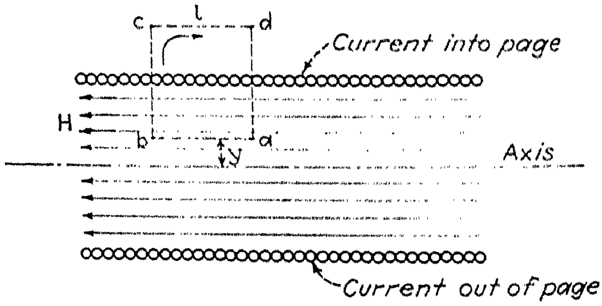

FIG. 64.

of the circuital law. Figure 64 shows a portion of a section of the solenoid through its axis. From the symmetry we see that the lines of H and B must be straight lines parallel to the solenoid axis in the space inside the solenoid. (These lines will of course come out of the ends and close on themselves. If the solenoid is long enough compared to its radius, however, the lines of H or B outside the solenoid will be far from it.) H can depend only on the distance y from the axis. Let us evaluate the magnetomotive force around the dotted path. For the portion of the path lying outside the solenoid, H is practically zero, and the portions normal to the axis give no contribution to the m.m.f. since H is at right angles to the path. Thus the whole integral takes its value from the portion ab of the closed path of length l, and we have

$$\oint H_s \, ds = Hl$$

where H is the magnitude of the magnetic intensity at a distance y from the axis. The current flowing through the area enclosed by this path is nli, where n is the number of turns per unit length and i is the current flowing in each turn. Thus we have

$$\oint H_s\, ds = Hl = 4\pi(nil)$$

so that

$$H = 4\pi ni \tag{41}$$

and is constant, independent of position inside the solenoid. This result checks Eq. (36) for the field on the axis of such a solenoid, but we now see that the result holds for points other than those on the axis. The corresponding value of B is

$$B = \mu_0 4\pi ni \tag{42}$$

for vacuum, and here we have an example of a uniform field of magnetic induction. In practice one never has an infinite solenoid, but, if the diameter of the solenoid is very small compared to its length, Eqs. (41) and (42) are correct for the central

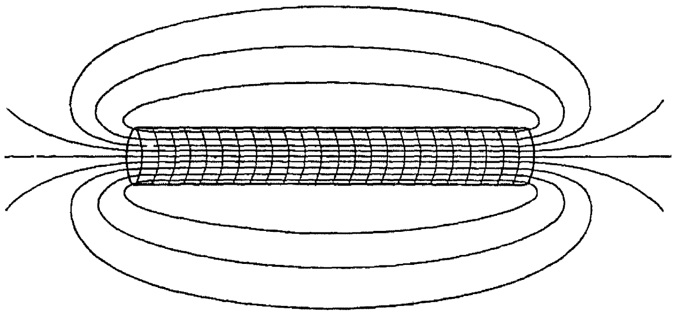

FIG. 65.

portion of the solenoid. Only near the ends where the lines start to curve do these equations give incorrect values of the field. Figure 65 gives a rough picture of the complete field of a solenoid of finite length.

31. Magnetic Moment of a Current Loop; Scalar Magnetic Potential.—Let us return to the problem of the magnetic field produced by a circular current-carrying loop. In Sec. 29 we derived the expression (Eq. 34) for the magnetic intensity H at any point on the axis of such a loop

$$H = \frac{2\pi i r^2}{(x^2 + r^2)^{\frac{3}{2}}}$$

where r is the radius of the loop and x the distance from the center of the loop to the point on the axis where H is calculated. If the radius r of the loop is very small compared with x, the above expression becomes, very nearly,

$$H = \frac{2\pi i r^2}{x^3} = \frac{2iA}{x^3} \tag{43}$$

where A is the area of the loop.

This expression is just like the corresponding expression for the electrostatic field on the axis of a dipole at distances large compared with the dimensions of the dipole. The correspondence between the magnetic intensity at large distances from a small current loop and the electric field intensity or displacement at large distances from a small dipole is also true for points not on the axis. Therefore it is very convenient to describe the magnetic field of a small current loop in terms of an equivalent *magnetic dipole*. For an electric dipole the expression for D, let us say, corresponding to Eq. (43) is

$$D = \frac{2p}{x^3} \tag{43a}$$

where p is the dipole moment, and D is in the same direction as the vector $\vec{p}$, from the negative to the positive charge of the dipole. We now *define* the magnetic moment $\vec{m}$ of a small loop of current as a vector of magnitude iA (product of current and area of the loop) whose direction is along the normal to the plane of the loop.

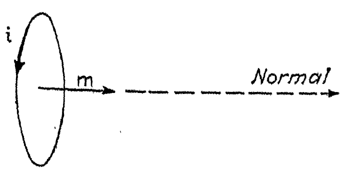

FIG. 66.

The sense of the vector $\vec{m}$ is such that if a right-handed screw were rotated in the direction of current flow it would move in the direction of $\vec{m}$ (Fig. 66). In terms of the magnetic moment of the current loop Eq. (43) takes the form

$$H = \frac{2m}{x^3} \tag{43b}$$

It must be always kept in mind that this expression for the axial field is valid *only* at points x which are very far from the loop.

Furthermore, in the electrostatic case the dipole field at large distances can be obtained by forming the gradient of the scalar potential V given by Eq. (25), Chap. II. The convenience of having a scalar potential has also been carried over to the case of a magnetic dipole, and we define the *scalar magnetic potential* of a dipole at large distances from it by (Fig. 67)

$$(44)$$

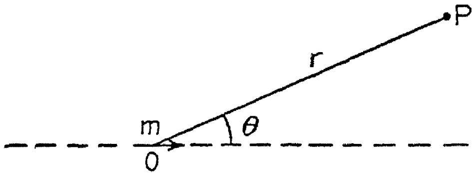

FIG. 67.

and from this the magnetic intensity H can be obtained as

$$H = - \text{grad } V, \tag{45}$$

It turns out experimentally that the magnetic field produced by a tiny bar magnet is identical at large distances with that produced by a tiny current loop; hence one thinks of such a permanent magnet as being describable by a magnetic moment. This equivalence is valid also for large bar magnets and solenoids, *provided* one restricts one's attention to the field *outside* the bar magnet or solenoid. In fact, from measurements of the external field produced by a bar magnet or by a solenoid, one could not distinguish between them. The scalar magnetic potential is particularly convenient for describing the external magnetic field of permanent magnets. One more word is necessary concerning the use of a scalar potential in magnetic calculations. The condition for the existence of such a quantity is that the magnetomotive force around a closed path be zero ($\oint H_s\, ds = 0$). We have seen that this is true for paths not enclosing a current. If, however, we consider a region of space in which distributed currents flow, the above equation is not true, and one *cannot* use the idea of scalar potential at all.

Equation (44) for the scalar magnetic potential of an infinitesimal current loop can be put in an interesting and useful form. Writing it as

$$V \qquad i\, dA \cos\theta$$

we see from Fig. 68 that the tiny area dA' normal to r is related to dA by the equation

$$dA' = dA \cos\theta$$

and that the solid angle $d\Omega$ subtended at P by dA is

$$d\Omega = \frac{dA'}{r^2} = \frac{dA \cos\theta}{r^2}$$

Hence the potential V_m can be written very simply as

$$V_m = i\,d\Omega \tag{46}$$

Written in this form one can now deduce the corresponding expression for a finite current loop of any shape whatsoever. If we have a large current-carrying loop, we can imagine it built up by superposing a huge number of tiny loops, each carrying the same current, as shown in Fig. 69. Any part of a loop not on the perimeter carries two equal and opposite currents (hence

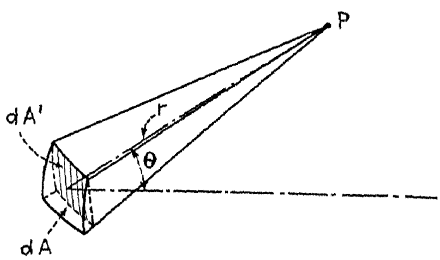

Fig. 68.

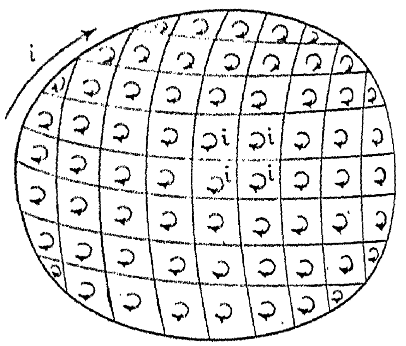

Fig. 69.

zero current), so that the large single loop is equivalent to the sum of the small ones. This construction bears the name of Ampère. At a point P the large loop subtends a solid angle Ω which is evidently the sum of the solid angles subtended by the tiny loops. Hence, using Eq. (46), the potential due to any loop of current at a point P is

$$V_m = i\Omega \tag{47}$$

where i is the current in the loop and Ω the solid angle subtended by the loop at P.

As an illustration of the use of Eq. (47) we calculate the scalar magnetic potential of a circular loop of radius r at a point on the axis of the loop. The solid angle Ω subtended at P by the loop is equal, by definition, to the area of the portion A' of the surface

of a sphere of radius R divided by R^2 (Fig. 70). The area A' is readily found to be

$$A' = 2\pi R^2(1 - \cos\theta)$$

where θ is the angle between the radius to a point on the loop and the x-axis as shown.

Thus

$$\Omega = \frac{A'}{R^2} = 2\pi(1 - \cos\theta) = 2\pi\left(1 - \frac{x}{R}\right) = 2\pi\left(1 - \frac{x}{\sqrt{x^2 + r^2}}\right)$$

and from Eq. (47),

$$V_m = 2\pi i\left(1 - \frac{x}{\sqrt{x^2 + r^2}}\right) \tag{48}$$

This gives the magnetic scalar potential at points on the x-axis and we can obtain the axial field intensity H according to Eq. (45) by

$$H = -\frac{dV_m}{dx} = 2\pi i\left[\frac{1}{\sqrt{x^2 + r^2}} - \frac{x^2}{(x^2 + r^2)^{\frac{3}{2}}}\right] = \frac{2\pi i r^2}{(x^2 + r^2)^{\frac{3}{2}}}$$

and this is the result expressed by Eq. (34).

It is interesting and instructive to express the torque on a current loop which is placed in a magnetic field in terms of the magnetic moment of the current loop. We have seen in Sec. 27 that a current-carrying loop in a uniform field B is in equilibrium when the plane of the loop is at right angles to the direction of B. From an examination of Fig. 54 on page 93, we see that the position of stable equilibrium is such that the magnetic field *produced* by the current in the loop *aids* the external field at the center of the loop. If the normal to the plane of the coil makes an angle θ with the direction of B, the restoring torque is

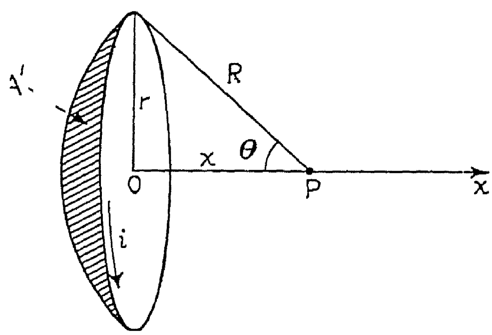

FIG. 70.

$$T = BiA\sin\theta \tag{49}$$

If the induction B varies from point to
equation is still correct for a very small loop, so small

that B is practically constant over the area A. In terms of the magnetic moment $m = iA$, Eq. (49) becomes

$$T = mB \sin \theta \tag{50}$$

and the axis of rotation (the direction of the vector torque) is normal to the plane of $\vec{m}$ and $\vec{B}$. Using the notation for a vector product, we can state this concisely as

$$\vec{T} = \vec{m} \times \vec{B} \tag{51}$$

(In Fig. 54*b* the direction of $\vec{m}$ is that of the normal n shown there.) We have already pointed out the equivalence of a small bar magnet, such as a small compass needle, and a small current loop with regard to the field produced by each. This equivalence is also true with respect to the torques exerted on each by a magnetic field. The torque exerted on a small compass needle suspended at its center is given by Eq. (51), where $\vec{m}$ now represents the magnetic moment of the compass needle. If the needle is displaced from its equilibrium orientation (in which it is lined up with the field vector B) through a *small* angle θ and released, it will perform simple harmonic motion with a period

$$T_0 = 2\pi\sqrt{\frac{I}{mB}} \tag{52}$$

where I is the moment of inertia of the needle about the axis of rotation. Used in this manner, a small magnet serves as a magnetometer and is employed to measure weak magnetic fields, such as the magnetic field of the earth.

There is a simple important current-measuring instrument, which operates with the aid of such a suspended compass needle, called a tangent galvanometer. Suppose a compass needle is suspended so that it is free to rotate about a vertical axis through its center and is placed at the center of a large vertical circular coil of wire of N turns and radius a. The plane of the coil is perpendicular to the east-west direction. When no current flows through the coil, the compass needle points north. If a current i flows through the coil, it sets up a field B at its center in an east-west direction of magnitude given by (see Eq. 33)

$$B_c = \mu_0 \frac{2\pi N i}{a}$$

The compass needle rotates until it points in the direction of the resultant of this field and the horizontal component of the earth's field, B_h. The tangent of the angle between the direction of the needle and north is given by (Fig. 71)

$$\tan\theta = \frac{B_c}{B_h} = \frac{2\pi\mu_0 N}{aB_h}\cdot i$$

or

$$\tan\theta = \text{constant} \times i \qquad (53)$$

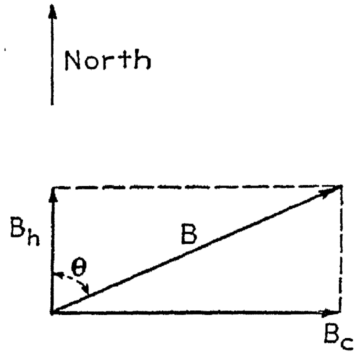

FIG. 71.

Thus we have a simple and relatively cheap ammeter which can be used to determine the absolute value of the current i if B_h is once known. If one were to mount a small current-carrying coil at the center of a large vertical circular coil so that the small coil were free to rotate about a vertical axis through its center, it would behave just like the compass needle. It is, of course, much more convenient to employ the small permanent magnet. Suppose the large coil has N turns and radius R and the small coil n turns and radius r. If we hold the small coil so that its plane is normal to that of the big coil and send the same current i through both coils (series connection), we must exert a torque on the small coil equal to

$$T = \mu_0 \frac{2\pi N i}{R}\cdot n i \pi r^2$$

or

$$T = \frac{2\pi^2\mu_0 N n r^2}{R}\cdot i^2 \qquad (54)$$

This equation can be used to define the abampere (the e.m.u. of current), placing $\mu_0 = 1$. Thus, for example, let $N = n = 10$, $r = 1$ cm. and $R = 100$ cm. Then the abampere is that current sent through the coils for which a torque of $2\pi^2$ dyne-cm. is exerted on the small coil.

In conclusion, we should point out that the definition of the magnetic moment of a very small current loop which we have adopted, viz., $m = iA$, is at variance with that adopted by many writers on the subject. The alternative definition, namely,

$m' = \mu_0 iA$ for a loop in vacuum, has enjoyed considerable popularity, and, using it, one is led to equations differing slightly from those we have written. For example, Eq. (44) for the magnetic scalar potential becomes

$$V_m = \frac{m' \cos \theta}{\mu_0 r^2}$$

and Eq. (51) for the torque on a current loop takes the form

$$\vec{T} = \vec{m}' \times \vec{H}$$

Since B is the fundamental magnetic force vector, this last relation seems forced and may lead to confusion when applied to permanent magnets. We return to a more critical discussion of this point when we study the magnetic behavior of ferromagnetic bodies.

Problems

1. The magnetic induction in a given region of space is given by $B_x = ay$, $B_y = -ax$, $B_z = 0$. Show that the lines of induction are circles.

2. In Prob. 8, Chap. I, the two deflecting plates of the cathode ray tube are replaced by coils producing a uniform transverse magnetic field of 5.3 gauss in the same region of space in which the electric field existed with the deflecting plates in position. The beam is deflected 3.1 cm. Calculate the ratio e/m for an electron.

3. An electron enters a uniform magnetic field perpendicular to the lines of induction and performs circular motion with a period of 10^{-8} sec.

a. Calculate the value of B in gauss and in webers per square meter.

b. If the electron enters with a speed acquired by falling from rest through a potential difference of 3,000 volts, what is the radius of the circular orbit?

4. Carry through the calculations of Prob. 3 for a hydrogen ion instead of an electron. The mass of a hydrogen ion is 1,840 times that of an electron.

5. The earth's magnetic field at the equator is about 0.4 gauss. What should the velocity of an electron be if it is to describe a circle around the earth at the equator? What is its energy in electron-volts? Should it travel toward the east or the west?

6. A magnetron consists of a filament 0.5 mm. in radius surrounded by a coaxial cylindrical plate of inner radius 5 cm. When the tube is placed in a uniform magnetic field with its axis parallel to the lines of B, it is observed that the electron current from filament to plate is zero for plate potentials less than 10.2 volts. Calculate the magnitude of B.

HINT: Set up the equation for conservation of mechanical energy, and set the torque about the axis equal to the rate of change of angular momentum.

7. Referring to Fig. 52 in the text, show that all electrons with the same speed v_0, leaving the slit S at small angles to the normal, will be brought to a focus at P.

8. Electrons are emitted normally from the negative plate of a parallel-plate condenser of separation d with negligible velocities under the action of ultraviolet light. The condenser is situated in a uniform magnetic field with the lines of B parallel to the plates, and a potential difference V is maintained between the plates. Show that no electrons reach the positive plate if $V < \frac{1}{2}\frac{e}{m}d^2B^2$.

HINT: Use Newton's second law for the motion parallel to the plates and the conservation of mechanical energy.

How should this inequality be written if the Gaussian system of units is to be employed?

9. A cloud chamber enables one to observe the tracks of charged particles. When a uniform magnetic field exists in the chamber, the tracks are curved. Suppose tracks of 102-mm. radius are observed when the field is 0.1 weber/meter^2 and that it is known that the particles have the same charge as an electron and an energy of 3,400 electron-volts. What is their mass?

10. A 60-mil diameter copper wire (density 8.9 grams/cm.^{3}) 1 ft. long is pivoted at A and rests lightly against a horizontal wire passing through B, as shown in Fig. 72. A current of 1 amp. is sent through the wires from A to C and a magnetic field of 1,000 gauss is applied normal to the plane of the wires. What angle θ will the hanging wire make with the vertical at equilibrium?

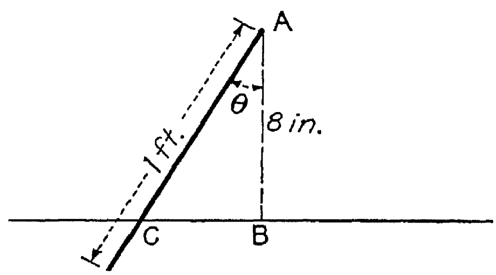

FIG. 72.

11. A copper bar weighing 100 grams rests on two rails 20 cm. apart and carries a current of 20 amp. from one rail to the other. The coefficient of friction is 0.16. What is the least magnetic field that would cause the bar to slide, and what is its direction?

12. A very thin wooden block, 25 by 40 cm., has 10 turns of wire wound around its edge, and the block is suspended in equilibrium with its plane horizontal on an east-west axis with its 40-cm. edges pointing north and south. Using very flexible leads, a current of 28 amp. is sent through the coil. How far from the axis of suspension must a 1-gram body be hung so that the coil remains in equilibrium in the earth's field? The horizontal component of the earth's magnetic field is 0.21 gauss, and its vertical component is 0.48 gauss.

13. An equilateral triangular loop of wire, of side a and weight W, is suspended from one vertex so as to turn freely in all directions. A current i flows around the loop, and it is placed in a uniform magnetic field B.

a. The field is normal to the loop. Draw the forces acting on the wires, and find the resultant tension in the wires.

b. The field is horizontal and in the plane of the loop. What is the torque tending to turn the loop?

c. The field is vertical. What is the equilibrium position of the loop?

14. A circular loop of wire of area A, carrying a current i, is placed in a uniform magnetic field B so that the field is in the plane of the loop. Show that the torque on the loop is BiA.

15. A uniform magnetic field is applied normal to a rigid circular loop of wire carrying a current. Find the tension in the wire.

16. The coil of a d'Arsonval galvanometer has 100 turns and encloses an area of 5 cm.² The magnetic induction in the region where the coil is located is 1,000 gauss and the torsional constant of the suspension is 10^{-5} dyne-cm./deg.

a. Find the angular deflection of the instrument per milliampere.

b. If the angular deflection is measured by reflecting a beam of light from a mirror on the suspension, what current would cause a deflection of 1 mm. of the light spot on a scale 1 meter distant from the mirror?

17. How will the sensitivity of a galvanometer change if the number of turns of wire on the coil is doubled, the diameter of the wire being correspondingly reduced so as to keep the weight constant? Consider both the ampere and the volt sensitivities.

18. Discuss how the sensitivity of a d'Arsonval galvanometer depends on the magnetic induction B, the area of the coil, and the length and radius of the suspending fiber. What practical limitations can you give for the design of the instrument, if one desires maximum sensitivity?

19. Calculate the period of a moving-coil galvanometer for torsional vibrations, neglecting friction. Derive a formula for the angular deflection of such an instrument in terms of the current, magnetic induction B, number of turns N, coil area A, the period calculated above, and the moment of inertia of the instrument about the axis of suspension.

20. A large current is sent through the coil of a d'Arsonval galvanometer for a very short time. Neglecting friction, prove that the maximum angle through which the coil rotates is proportional to the total charge passing through the coil.

Derive a formula for the proportionality constant. When a galvanometer is used in the manner described, it is called a ballistic galvanometer. Why should a ballistic galvanometer have a relatively large moment of inertia about the suspension axis?

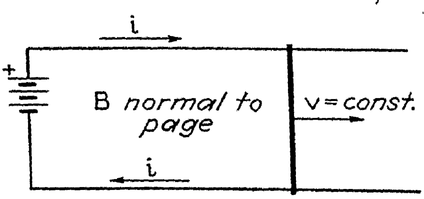

Fig. 73.

21. A metal bar slides on two conducting rails carrying a current i, as shown in Fig. 73. Under the influence of a uniform magnetic field B and a suitable external force, it moves with the velocity v. Compare the rate at which work is done against the external force with the rate of increase of flux in the circuit behind the wire. Give your answer for both m.k.s. and absolute electromagnetic units.

22. The armature of a motor is a cylinder 25 cm. in diameter and rotates with an angular velocity of 1,200 r.p.m. The armature carries 100 conduc-

tors, each 40 cm. long and each carrying 15 amp. at right angles to a magnetic field of flux density 8,500 gauss. Compute the torque and horsepower developed by the motor.

23. Two long straight parallel wires carry equal and opposite currents. The wires are separated by a distance of 10 cm. and each carries a current of 20 amp.

a. Calculate an expression for the magnetic intensity H at any point (outside the wires) in the plane containing the wires.

b. Plot to scale the magnitude of H as a function of distance along a line perpendicular to both wires, using the mid-point between the wires as an origin.

24. A long straight wire of circular cross section and radius 0.02 cm. carries a steady current of 50 amp. Calculate the total flux of B inside a cylinder coaxial with the wire of radius 2 cm. and altitude 1 meter. Neglect the flux *inside* the wire.

25. A solenoid of 1,000 turns is wound uniformly on a cylindrical tube 40 cm. long and of 8 cm. radius. A current of 1.5 amp. flows in the winding. Calculate the axial magnetic intensity at the center of the solenoid and in the plane at one end of the solenoid.

26. Derive a formula for the magnetic intensity at points on the axis of a solenoid of length l in terms of the distance from the center of the solenoid. Plot the magnitude of H against this distance for points both inside and outside the solenoid.

27. Calculate the attractive force per foot between two long parallel wires 5 in. apart when each carries a current of 50 amp. Express your answer in pounds per foot.

28. Derive an expression for the magnetic induction at the center of a square circuit of side l, if the circuit carries a current i.

29. Calculate the magnetic intensity due to two long parallel conductors separated by a distance $2d$ and carrying equal currents in opposite directions at any point of a plane bisecting a line joining the two wires and perpendicular to each. Where is this intensity a maximum?

30. Calculate and plot the magnetic induction as a function of distance from the central axis of a long cylindrical wire 2 mm. in radius carrying a current of 50 amp. What is the flux of B per centimeter of length *inside* the wire?

31. A steady current i flows in a long straight cylindrical wire of radius a and returns along a coaxial hollow cylinder of inner radius b and thickness d. Assuming uniform current density in the conductors, calculate the magnetic intensity as a function of distance from the central axis of the wire. (Use Ampère's circuital theorem.)

32. In Prob. 31 calculate the ratio of the flux of B outside the conductors to that inside both conductors.

33. A uniformly charged circular ring is rotated about its axis with constant angular velocity ω. Calculate the magnetic field intensity

a. At the center of the circle,

b. At any point on the axis.

34. A sphere carries a uniform surface-charge density σ and is r about a diameter with constant angular velocity ω. Using the results of Prob. 33, calculate the magnetic induction B at the center of the sphere.

35. A toroid of inner radius 9 cm. and outer radius 10 cm., as shown in Fig. 74, is wound uniformly with 2,000 turns of wire. If a current of 4 amp. is sent through the winding what is the magnetic induction at a point 9.2 cm. from the axis of the toroid? By how many per cent does B vary as one moves from a point 9 cm. from the axis to a point 10 cm. from it?

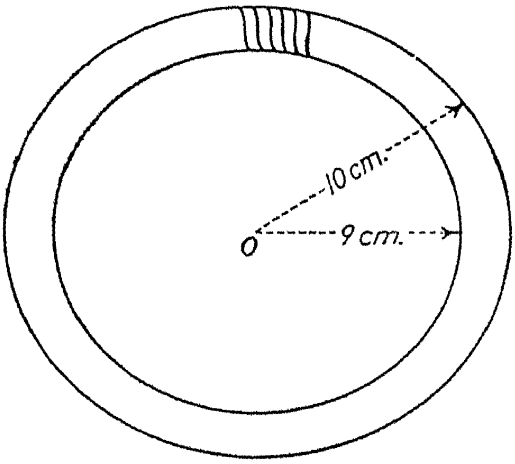

FIG. 74.

36. A brass rod of square cross section (2 × 2 cm.) is bent into the form of a ring of inner radius 5 cm., and the ends are welded together. Wire is wound toroidally around this ring to form a coil of 500 turns, and a current of 1 amp. flows through the winding.

Assuming the magnetic permeability of brass to be that of empty space, calculate the total magnetic flux (of B) in the brass.

Note that B varies from point to point inside the brass.

37. Calculate an expression for the rate of change of the axial magnetic field intensity with distance along the axis of a circular turn of radius a carrying a current i.

Show that the above expression has a zero rate of change at a point on the axis at a distance $a/2$ from the center of the turn.

What value has H at this point?

38. The results of Prob. 37 are used in the design of *Helmholtz coils* used to produce a uniform magnetic field in a small region of space. Two coils, each of N turns and radius a, are mounted coaxially and separated by a distance a. The field due to both coils is nearly uniform in the neighborhood of the point on the axis midway between them, *i.e.*, at a distance $a/2$ from either coil.

Calculate and plot the resultant axial field due to both coils as a function of position along the common axis of the coils. What is the value of H at the central point midway between the coils?

39. Two circular turns of wire, each 20 cm. in radius, are mounted coaxially at a separation of 20 cm. A current of 10 amp. flows steadily through each turn, and this device is used as a tangent galvanometer. A compass needle is mounted at a point on the axis midway between the turns, and the common axis lies along the east-west direction. If the horizontal component of the earth's magnetic field is 0.2 gauss, calculate the angle between the needle and the axis of the system.

40. Show by Ampère's law that the magnetic intensity on either side of a plane current sheet is $2\pi j$, where j is the surface density of current, and is directed parallel to the plane and at right angles to the current.

41. A long thin strip of metal, of width w, and negligible thickness, carries a current i. It is placed in a uniform magnetic field of intensity $2\pi i/w$ directed parallel to the plane of the strip and at right angles to the current. Show that the resultant field is zero on one side of the strip and $4\pi i/w$ on the other. Show that there is a force on the metal strip equivalent to a pressure $\mu_0 H^2/8\pi$ acting on the side where there is the field H (Fig. 75).

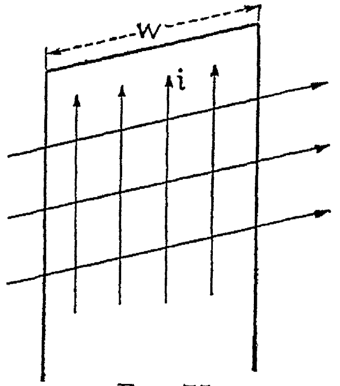

FIG. 75.

42. Coils are frequently wound in "pancake sections," each section being a closely wound plane spiral of wire. Find the force between two neighboring sections if they are wound of 3-mm. wire, if the inside diameter of the section is 5 cm., the outside diameter 15 cm., and if the wire carries 30 amp. Consider a section as equivalent to a current distribution in a plane, the lines of flow being circles, the current density being constant, and the lines of B being radial just outside the plane.

43. Find an expression for the resultant force between two infinite straight wires carrying equal current if the wires are not in the same plane and form an angle θ with each other.

44. A metal roller rests on two parallel conducting rails. Show that, when the rails are connected to a source of e.m.f. so that a steady current flows through the rails and roller, the roller tends to move.

CHAPTER VI

INDUCED ELECTROMOTIVE FORCES AND INDUCTANCE

In the preceding chapter we considered the magnetic fields produced by steady currents and the forces which these fields exerted on conductors carrying steady currents and on moving charges. The magnetic field produced by a distribution of steady currents is stationary, the value of B, for example, being constant at a given point of space. Stationary magnetic fields exert no forces on conductors at rest if the latter carry no current. If, however, the magnetic field in the neighborhood of a stationary conducting circuit changes with time, one observes a current flow in the circuit, a so-called induced current, and this current flows as long as the magnetic field keeps changing. One interprets this current flow as being caused by an electromotive force induced in the circuit, and these induced e.m.fs. were discovered by Faraday in 1831 and also independently by Henry.

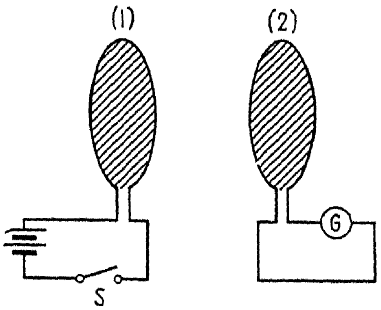

FIG. 76.

Suppose we have two fixed conducting loops arranged as in Fig. 76. If the switch S is closed, a momentary deflection of the galvanometer G is observed, ceasing as the current flow in circuit (1) becomes steady. If the switch is then opened, the galvanometer deflects in the opposite direction. Thus we see that, while the current in circuit (1) is increasing or decreasing, with a consequent increase or decrease of the magnetic field, there is an induced e.m.f. in circuit (2). Faraday also discovered that if circuit (2) is moved relative to circuit (1), while the latter carries a *steady* current, an e.m.f. is also induced in circuit (2). This motion may be such that the circuit (2) remains rigidly undeformed or may consist of a deformation of the circuit. The e.m.fs. induced in conductors which move relative to a stationary magnetic field are often called *motional* e.m.fs., and these are the

electromotive forces induced in many types of rotating machinery. In the following sections we shall formulate the laws governing these induced voltages and examine some of their consequences.

32. The Faraday Induction Law for Stationary Circuits; Lenz's Law.—Let us start with a discussion of the e.m.f. induced in a stationary circuit, which we take for simplicity as a single circular turn of wire. If the magnetic field in the region of space occupied by the turn is changed in any manner, *e.g.*, by changing the current in neighboring circuits, or moving the latter when they carry steady currents or by moving a permanent magnet in the neighborhood, *the induced electromotive force is found to be equal to the rate at which the magnetic flux which crosses any area of which the loop is a boundary changes with time.* The magnetic flux crossing such a surface is said to "link" the circuit.

Denoting the induced e.m.f. by $E = \oint \mathcal{E}_s \, ds$ and the flux linking the circuit by $\Phi = \int B_n \, dS$, the Faraday induction law becomes

$$E = -\frac{d\Phi}{dt} \tag{1}$$

The negative algebraic sign occurring in Eq. (1) is to indicate the direction of the induced e.m.f. and hence that of the induced current. The induced current always flows in such a direction that its magnetic field tends to *oppose the change* in the magnetic field which produces it. Thus, if we try to increase the magnetic flux through the loop ($d\Phi/dt$ positive), the induced current flows so that its field tends to decrease the flux and opposes the *change.* Hence the minus sign indicates the induced voltage establishing a current tending to decrease the flux in the above example.

This law concerning the direction of induced e.m.fs. is perfectly general, holding equally well when the e.m.fs. are induced in moving circuits, and it bears the name of *Lenz's Law.* It follows immediately from the conservation of energy since otherwise an induced e.m.f. once started would grow indefinitely large. If, instead of a single conductor loop, one considers a coil of N turns in series, Eq. (1) becomes

$$E = -\sum_{k=1}^{N} \frac{d\Phi_k}{dt} \tag{2}$$

where Φ_k is the flux linking the kth turn. An important special case occurs when all the Φ_k's are equal (or very nearly so), the same flux linking each turn. Eq. (2) then becomes

$$E = -N\frac{d\Phi}{dt} = -\frac{d(N\Phi)}{dt} \tag{3}$$

and the product $N\Phi$ is called the number of flux linkages. Equation (1), (2), or (3) may be used as it stands in the m.k.s. or electromagnetic systems of units. In the former case, Φ is measured in webers, t in seconds, and E comes out in volts. In the latter case, Φ is expressed in maxwells, t in seconds, and E in abvolts (the e.m.u. of e.m.f.). It is convenient to remember that 1 *volt* = 10^8 *abvolts.*

If one wishes to employ the Gaussian mixed system, expressing Φ in maxwells, t in seconds, and E in statvolts, the equations *cannot* be used as they stand. The correct form is readily shown to be [for Eq. (1), for example]

$$E = -\frac{1}{c}\frac{d\Phi}{dt} \tag{3a}$$

where c is 3×10^{10} cm./sec. The other two equations are similarly altered.

The Faraday induction law holds equally well for extended conducting bodies at rest. In this case it is more convenient to write Eq. (1) in the equivalent form

$$\oint \mathcal{E}_s\, ds = -\int \frac{dB_n}{dt} dS = -\frac{d}{dt}\int B_n\, dS \tag{4}$$

Here the e.m.f. is taken around *any* closed path in the conducting medium, and the integral $\int B_n\, dS$ is the flux linking this path. The spacially distributed currents which flow under such conditions are called *eddy* currents. There is a further generalization of the induction law which we shall need later. *We postulate that an electric field is established in any region of space in which a magnetic field changes with time and that Eq.* (4) *is obeyed whether the closed path be in a conducting or nonconducting medium or in empty space.* In this connection it must be emphasized that the electric field so produced is quite different in nature from that produced by stationary charges. In fact, Eq. (4) shows clearly that $\oint \mathcal{E}_s\, ds$ is not zero, and hence that no scalar potential for

$\mathcal{E}$ exists. One can no longer obtain $\mathcal{E}$ as the negative gradient of a scalar potential when discussing the electric field produced by changing magnetic fields. This is, of course, in sharp contrast to the electrostatic case.

33. Motional Electromotive Forces.—We now turn our attention to the case of induced e.m.fs. caused by the motion of conductors relative to a magnetic field. An exact analysis of the situation requires relativity theory which is far beyond the scope of this book; hence we must content ourselves with an approximate method which gives excellent results provided the velocities encountered are small compared to c (3×10^{10} cm./sec). For most practical problems this condition is well satisfied. We shall approach the question from the law giving the force exerted by a magnetic field on a moving charge, the so-called *Lorentz* force. This force is

$$F = q(\overrightarrow{v} \times \overrightarrow{B})$$

where $\overrightarrow{v}$ is the vector velocity of the charge q. If we now imagine a metallic circuit moving in a fixed magnetic field, the conduction electrons in it will be acted on by a force per unit charge equal to $(\overrightarrow{v} \times \overrightarrow{B})$ and, being free to move, will set up an induced current. Treating this force per unit charge as an effective electric intensity $\mathcal{E}'$ from the standpoint of an observer moving with the circuit, we expect that the magnitude of the induced e.m.f. should be given by a formula

$$E = \oint \mathcal{E}'_s \, ds = \oint (\overrightarrow{v} \times \overrightarrow{B})_s \, ds \tag{5}$$

Equation (5) is the correct formula for low velocities, and the induced e.m.f. as given by this equation sets up a current which, in accordance with Lenz's law, tends to oppose the motion producing it. There will be a force exerted on the conductors tending to slow them down, and this force action is usually termed *electromagnetic reaction.* In many applications of Eq. (5) it is convenient to think of an e.m.f. induced in each element ds of the circuit

$$dE = \mathcal{E}'_s \, ds = (\overrightarrow{v} \times \overrightarrow{B})_s \, ds \tag{6}$$

and the e.m.f. along any path becomes the sum of these infinitesimal e.m.fs. Viewed in this manner, one can state that *the*

e.m.f. induced in a conductor is equal to the rate at which this conductor cuts lines of B. We can see this readily with the help of a simple example. Suppose a metal bar is moved at constant velocity in a direction perpendicular to its length in a uniform field of magnetic induction directed normal to the conductor and to its direction of motion, as shown in Fig. 77. In time dt the conductor sweeps out an area equal to $lv\,dt$ and cuts all the lines of B which cross this area. The number of lines cut is

$$d\Phi = Blv\,dt$$

so that the induced e.m.f. has a magnitude given by

$$E = \frac{d\Phi}{dt} = Blv \tag{7}$$

This is exactly the result which one obtains employing Eq. (6), remembering that v is normal to B and that $v \times B$ is directed along the length of the conductor. Furthermore, v and B have the same values and directions at each point of the conductor. Other cases will be taken up in the problems.

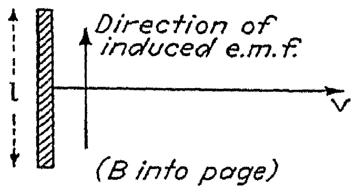

FIG. 77.

If we apply Eq. (5) to a moving closed circuit or to a closed circuit in an extended conductor in motion, we can, under certain conditions, find a simple relation between the rate of change of flux through the moving circuit and the induced e.m.f. as given by this equation. *If the circuit moves as a rigid body,* undeformed during its motion, *the induced e.m.f. as given by Eq.* (5) *is just equal to the rate of change of flux through the circuit.* This is true if the magnetic field is stationary. If the magnetic field varies with time, one gets the correct induced e.m.f. by taking the *total* rate of change of flux through the circuit. This now consists of the sum of two terms, one the rate of change of flux due to the varying field and the other the rate of change of flux due to the motion of the circuit in the field.

We shall not attempt a general proof of the above statements but shall content ourselves with a simple, but important, example. Consider a rigid rectangular coil of width d and length l which is rotated with constant angular velocity ω about an axis perpendicular to a uniform magnetic field B, as shown in Fig. 78. The angle θ between the direction of B and that of the normal

n to the plane of the loop varies uniformly with time, so that $\theta = \omega t$. Let us first evaluate the induced e.m.f. in the coil by applying Eq. (5). We note that the vector $\vec{v} \times \vec{B}$ is directed parallel to the axis of rotation for all elements ds of the coil, and hence the induced e.m.f. may be considered to arise in the sides l only. This is in accord with the notion of cutting lines of B, and the two e.m.fs. thus induced add as one moves around the loop. For each side l we have

$$(\vec{v} \times \vec{B}) = vB \sin \theta$$

and this vector is along the direction l; hence we have

$$E = \oint (\vec{v} \times \vec{B})_s \, ds = 2vBl \sin \theta = 2vBl \sin \omega t$$

The speed v of either side l is related to the angular velocity ω by $v = \omega d/2$, so that the induced e.m.f. becomes

$$E = \omega Bld \sin \theta = \omega BA \sin \omega t \quad (8)$$

Now let us calculate this e.m.f. by evaluating the rate of change of flux through the coil. When in the position shown in Fig. 78, the flux Φ linking the turn is

$$\Phi = \int B_n \, dS = BA \cos \theta = BA \cos \omega t$$

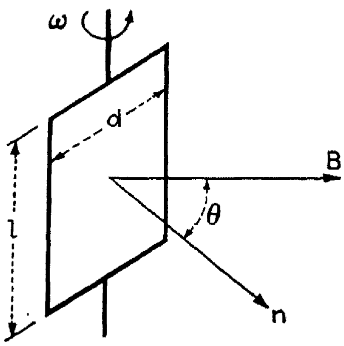

FIG. 78.

and differentiating with respect to the time t, we find

$$E = \omega BA \sin \omega t$$

which is identical with the result expressed by Eq. (8). Writing Φ_m for the maximum flux linking the coil and assuming that we have a coil of N turns in series, we have an induced voltage in the coil equal to

$$E = N\omega\Phi_m \sin \omega t \quad (9)$$

where $\Phi_m = BA$. Equation (9) is the fundamental formula underlying the action of either a.c. or d.c. generators. If the terminals of the coil are brought out to slip rings on the axis of rotation, we have the case of an a.c. generator in which the induced e.m.f. varies sinusoidally with the time. In the case of

d.c. generators a commutator, which reverses the direction of current flow in opposite sides l (with respect to a fixed external circuit) every half cycle, takes the place of the slip rings.

The case of induced voltages in circuits which are deformed and do not move as rigid bodies is more complicated, and in general it is safest to apply Eq. (5). For such examples it is true only in special cases that the induced e.m.f. is equal to the rate of change of flux through the circuit. One such special case is shown in Fig. 79, in which a metal rod ab, of length l, is moved with constant velocity v and slides along conducting rails ad and bc. The induced voltage around the circuit $adcb$ is $E = Blv$, and this is equal in magnitude to the rate of increase of flux through the circuit $adcb$. Since B is assumed uniform, the rate of increase of flux is equal to the flux density B times the rate of increase of area. This latter term is lv. The induced current, which flows as in the direction shown, has a magnitude

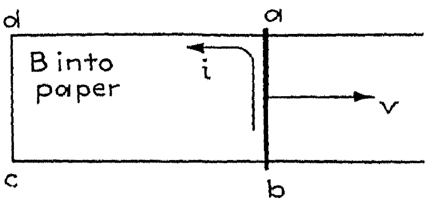

FIG. 79.

$$i = \frac{E}{R} = \frac{Blv}{R} \tag{10}$$

where R is the resistance of the circuit. Since the conductor ab carries this current, there is a side thrust on it equal to $F = Bli = B^2l^2v/R$, which is directed opposite to the direction of v. This is the electromagnetic reaction and one must exert a force equal and opposite to it on the conductor to keep it moving with a constant velocity. The power required is

$$P = Fv = \frac{B^2l^2v^2}{R} = \frac{E^2}{R} = i^2R \tag{11}$$

and we see that the law of conservation of energy is satisfied, the mechanical work done per unit time by external forces just equal to the rate of heating in the circuit. When one moves an extended metallic conductor in a magnetic field, there is an electromagnetic reaction similar to that discussed in the above example, and the conductor moves as if in a viscous medium, the force being proportional to the velocity and opposite to the direction of motion. This is the principle of the eddy-current brake.

In all the applications with which we shall concern ourselves, we shall restrict ourselves either to stationary circuits in varying magnetic fields or to moving circuits which are *not* deformed during the motion. In both these cases we may use the Faraday induction law as expressed by Eq. (1) to calculate induced e.m.fs.

34. Self- and Mutual Inductance.—In this section we shall restrict our attention to fixed, rigid circuits and shall examine the induced voltages created in them when the currents they carry change with time. First we consider a simple circuit as shown in Fig. 80, carrying a steady current i. There is a steady magnetic field set up by the current, and some of the lines of B are indicated. There is a definite flux linking the coil and the number of flux linkages with the coil is

$$= N\int B_n \, dS$$

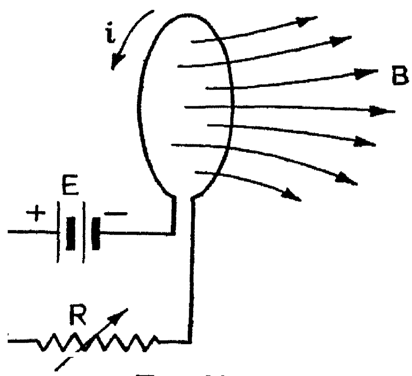

FIG. 80.

where N is the number of turns in the coil and the integral extends over an area bounded by the loop. The exact evaluation of this integral is exceedingly difficult in all but a few simple geometrical shapes of the circuit, but no matter what the geometry, the results of the preceding chapter require that it be proportional to the steady current i flowing in the loop (H and hence B at any point are proportional to i). If we denote the proportionality factor by L, we can write for the flux linkages

$$= Li \tag{12}$$

where L, called the self-inductance of the circuit, depends only on the geometry of the circuit and not on the current, at least in empty space.

If we now vary the current in the circuit by changing the variable resistance and if we *assume that, while the current is changing,* $N\Phi$ *is proportional to the instantaneous value of the current,* an induced e.m.f. will be generated in the circuit according to Eq. (1) whose magnitude is

$$= N\frac{d\Phi}{dt} = L\frac{di}{dt} \tag{13}$$

and which acts in such a direction as to oppose the change of current by Lenz's law. This is often called a back e.m.f. Our assumption concerning the validity of the proportionality $\Phi \sim i$ turns out to be very accurately true for slow rates of change of i with time, and for the present we shall accept it as true for all the cases we shall examine. A more exact criterion for its validity will be given in Chap. VIII.

According to Eq. (12) the self-inductance L of a circuit equals the number of flux linkages set up in the circuit per unit current in it. In the m.k.s. system the unit inductance is called the *henry* and this is 1 weber/amp. or, as we see from Eq. (13), 1 volt-sec./amp. In e.m.u., the unit inductance is 1 abhenry which is 1 maxwell/abampere or 1 abvolt-sec./abampere. *One henry* $= 10^9$ *abhenrys.*

In general, whenever the current in a circuit is caused to change, there will be induced in it a back e.m.f. given by Eq. (13) which tends to keep the current from changing. Thus, when one closes a switch in a circuit fed by a d.c. seat of e.m.f., the current will attain its final steady-state value only after some time. While the current is increasing, the net e.m.f. acting around the circuit is $E - L\dfrac{di}{dt}$, where E is the external e.m.f. Similarly a current cannot be instantaneously reduced to zero by opening a switch.

We now turn to a calculation of the inductance of two readily calculable circuits. First, consider a solenoid whose length l is large compared to its radius, n turns per unit length, N total turns carrying a steady current i. From Eqs. (41) and (42) of Chap. V we write for H and B at all points inside the solenoid

$$H = 4\pi ni$$

and

$$B = \mu_0 4\pi ni$$

These formulas are, strictly speaking, valid only for an infinitely long solenoid but are good approximations when l is large compared to the radius. This is equivalent to neglecting end effects in a manner similar to the treatment of a parallel-plate condenser in electrostatics. The total flux Φ inside the solenoid is accordingly

$$\Phi = BA = \pi r^2 B$$

where r is the radius and the total number of flux linkages is

$$N\Phi = \pi r^2 N\left(\frac{4\pi\mu_0 Ni}{l}\right)$$

where we have written

$$n = \frac{N}{l}$$

Thus we obtain from Eq. (12) a formula for the inductance L of a long solenoid in air;

$$L = \frac{N\Phi}{i} = \frac{4\pi^2\mu_0 N^2 r^2}{l} \tag{14}$$

or, written in terms of the number of turns per unit length,

$$L = 4\pi^2\mu_0 n^2 r^2 l \tag{14a}$$

It is evident from Eq. (14) that the dimensions of inductance in e.m.u. are those of length so that 1 abhenry equals 1 cm.

As a second example, consider a long coaxial cable consisting of a central cylindrical wire of radius a and an outer thin hollow cylinder of radius b (Fig. 81). We shall assume that $b - a >> a$ so that we can neglect the magnetic flux *inside* the conductors. Since the magnetic induction has the value

$$B = \frac{2\mu_0 i}{r}$$

at points in the space between the conductors and is zero outside, the total flux linking a length l of the circuit can be obtained by a simple integration. Consider an element of area $l\,dr$ as shown in the figure. The flux across the area is evidently

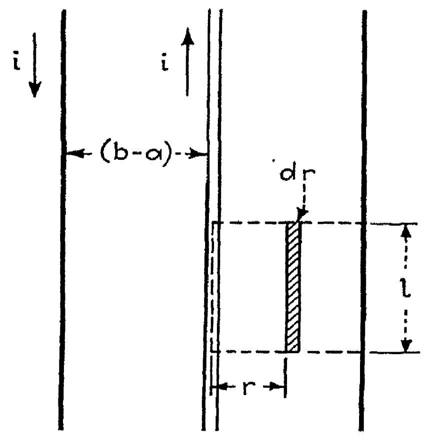

FIG. 81.

$$d\Phi = Bl\,dr = 2\mu_0 il\,\frac{dr}{r}$$

and the total flux linking a length l is

$$\Phi = 2\mu_0 il\int_a^b \frac{dr}{r} = 2\mu_0 li \ln\left(\frac{b}{a}\right) \tag{15}$$

This gives for the *inductance per unit length* of the cable

$$L' = \frac{\Phi}{il} = 2\mu_0 \ln\left(\frac{b}{a}\right) \tag{16}$$

Mutual Inductance.—Now let us consider the case of two circuits such as those shown in Fig. 82. If the current in circuit I is changed, a voltage is induced in circuit II and conversely, a current change in circuit II induces a voltage in circuit I. First take the case of a steady current i_1 in circuit I. This sets up a magnetic field as indicated in the figure, and a fraction of the flux due to i_1 links circuit II. Let the flux linking circuit II due to i_1 be denoted by Φ_{21}. This is proportional to N_1, the number of turns in circuit I, to i_1 and the proportionality constant depends only on the geometry of the system, *i.e.*, the dimensions and shapes of both circuits and their relative positions. In symbols, this is

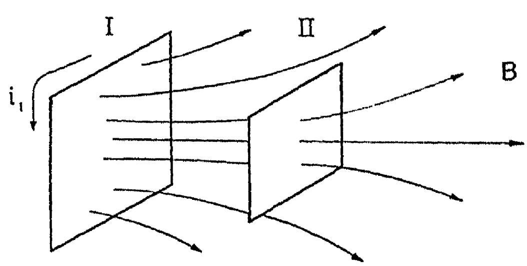

FIG. 82.

$$\Phi_{21} = KN_1i_1 \tag{17}$$

Now consider a steady current i_2 in circuit II and let us call the flux linking circuit I due to this current Φ_{12}. This flux is proportional to the number of turns N_2 of circuit II and to i_2. The proportionality constant depends only on the dimensions and shapes of the two circuits and on their relative positions and it can be shown to be the same as K in the above formula. Thus we can write

$$\Phi_{12} = KN_2i_2 \tag{18}$$

We now define the *coefficient of mutual inductance* M of the system as the number of flux linkages of either circuit per unit current in the other. Thus we have

$$M = \frac{N_2\Phi_{21}}{i_1} = \frac{N_1\Phi_{12}}{i_2} = KN_1N_2 \tag{19}$$

Now suppose we change the current i_1 in circuit I. This will induce an e.m.f. E_2 in circuit II given by

$$E_2 = -N_2\frac{d\Phi_{21}}{dt} = -M\frac{di_1}{dt} \tag{20}$$

If the current i_2 in circuit II is changed, an e.m.f. E_1 is induced in circuit I given by

$$E_1 = -N_1\frac{d\Phi_{12}}{dt} = -M\frac{di_2}{dt} \tag{21}$$

Any two circuits which are so arranged that an appreciable amount of flux due to one of them links the other are said to be *coupled*, and any change of current in one induces a voltage in the other. The maximum mutual inductance is obtained when *all* the flux produced by either circuit links the other.

Let us calculate the mutual inductance of two long coaxial solenoids. Suppose the length of each is l, the number of turns N_1 and N_2, and the radii r_1 and r_2, respectively. For the sake of definiteness, let us assume that $r_2 < r_1$. The induction B produced by a current i_1 in the outer solenoid is

$$B = \mu_0\frac{4\pi N_1 i_1}{l}$$

and the flux linking the inner solenoid is

$$\Phi_{21} = B\pi r_2^2 = \mu_0\frac{4\pi N_1 i_1}{l}\cdot\pi r_2^2$$

The mutual inductance is, according to Eq. (19),

$$M = \mu_0\frac{4\pi N_1 N_2}{l}\cdot\pi r_2^2 \tag{22}$$

The same result can be obtained by considering a current in the inner solenoid and calculating the number of flux linkages of the outer solenoid. This is left as an exercise for the student. With the help of Eq. (14), we can write for the self-inductances of the two solenoids

$$L_1 = \mu_0\frac{4\pi N_1^2}{l}\cdot\pi r_1^2 \tag{23}$$

and

$$L_2 = \mu_0\frac{4\pi N_2^2}{l}\cdot\pi r_2^2 \tag{24}$$

Comparing Eqs. (22), (23), and (24) one finds readily that

$$M = \frac{r_2}{r_1}\sqrt{L_1L_2} \tag{25}$$

This is a special case of the general relation between the mutual inductance and self-inductances for a coupled circuit. In general, one has

$$M = k\sqrt{L_1 L_2} \tag{26}$$

where k is called the *coefficient of coupling* and is always less than or at most equal to unity. If $k = 1$, $M = \sqrt{L_1 L_2}$ is the maximum mutual inductance possible. For the case of the two long solenoids we see that this occurs for $r_1 = r_2$, in which case all the flux produced by one links the other.

35. Energy Stored in the Magnetic Field of an Inductance; Energy Density.—In setting up a steady distribution of electric currents a definite amount of work must be done by the external sources of e.m.f. against the induced voltages which exist while the currents are being built up. We think of the work as being stored in the magnetic field associated with the current, and it can be regained by allowing the currents to vanish. To investigate these relations consider the simple circuit of Fig. 83, formed by connecting a coil of inductance L and resistance R to a battery. If the switch S is closed, the current i in the circuit will start to increase with a consequent increase of the magnetic field produced by this current. At any instant of time we have, from the Faraday induction law, applied to the circuit

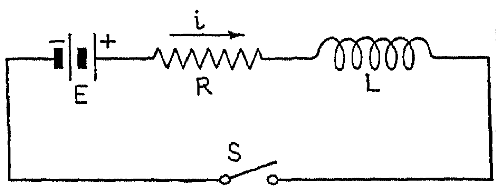

FIG. 83.

$$\oint \mathcal{E}_s \, ds = -N\frac{d\Phi}{dt} = -L\frac{di}{dt}$$

If we assume Ohm's law to be valid for currents varying with time, the left-hand side of this equation has the value

$$\oint \mathcal{E}_s \, ds = -E + iR$$

just as in the case of steady currents. E is the e.m.f. of the battery. Thus we have

$$E = iR + L\frac{di}{dt} \tag{27}$$

as the differential equation for the current in such a circuit. The work done per unit time by the battery, *i.e.*, the power input

to the system resistance plus inductance, is Ei. Using Eq. (27), we have

$$Ei = i^2R + Li\frac{di}{dt} = i^2R + \frac{d}{dt}\left(\frac{1}{2}Li^2\right) \tag{28}$$

The first term on the right-hand side is the rate of heating in the resistance, and we interpret the second as the rate of increase of magnetic energy. The total energy stored in the magnetic field produced by the inductance L after time t, when a current i is flowing, is accordingly

$$U_m = \int_0^t \frac{d}{dt}\left(\frac{1}{2}Li^2\right) dt = \int_0^i d\left(\frac{1}{2}Li^2\right) = \frac{1}{2}Li^2 \tag{29}$$

Just as in the case of electrostatic energy, we can think of this energy as distributed throughout the region of space in which the magnetic field exists and introduce the concept of *magnetic energy density*, the energy per unit volume of space. We can obtain a formula for this energy density by considering the magnetic field produced by an infinitely long solenoid carrying a steady current i. Consider a section of the solenoid of length l. Its inductance is, according to Eq. (14*a*),

$$L = 4\pi\mu_0 n^2 Al$$

where n is the number of turns per unit length and A the cross section of the solenoid. The current is related to the magnetic induction B inside the solenoid by

$$B = 4\pi n\mu_0 i$$

so that

$$i^2 = \frac{B^2}{16\pi^2 n^2 \mu_0^2}$$

Hence the magnetic energy $\frac{1}{2}Li^2$ is

$$U_m = \frac{1}{2}4\pi\mu_0 n^2 Al \cdot \frac{B^2}{16\pi^2 n^2\mu_0^2} = \frac{B^2}{8\pi\mu_0}Al$$

Now Al is just the volume of that region of space enclosed by a length l of the solenoid, so that the energy per unit volume of the space in which the magnetic field exists is given by

$$u_m = \frac{B^2}{8\pi\mu_0} = \frac{\mu_0 H^2}{8\pi} \tag{30}$$

This turns out to be the correct formula in general and holds even when B varies from point to point of space. From our derivation it is clear that this expression represents the work done against the induced e.m.fs. in building up the field with the help of currents. The total magnetic energy can be calculated from

$$U_m = \int_{\substack{\text{all}\\ \text{space}}} \frac{1}{8\pi\mu_0} B^2\, dv \tag{31}$$

and this can be evaluated if the space dependence of B is known

Problems

1. The magnetic flux linking a fixed coil of 1,000 turns is varied so that it has the value at any time t

$$\Phi = \Phi_m \sin$$

where $\Phi_m = 3{,}000$ maxwells and t is in seconds. Calculate an expression for the induced electromotive force E in the coil. What is the maximum value of this e.m.f. in volts? Make a plot of Φ and E as functions of time, plotting them on the same graph.

2. A slender metallic rod of length $2l$ rotates about an axis through its center with an angular velocity of n r.p.s. If the rod is placed in a uniform magnetic field B directed parallel to the axis of rotation, show that the magnitude of the voltage induced between the axis and either end of the rod is given by

$$E =$$

3. A copper disk of radius 10 cm. is rotated about its axis of symmetry at an angular velocity of 3,600 r.p.m. The disk is in a uniform field of magnetic induction of magnitude 200 gauss the direction of which makes an angle of 30° with the plane of the disk.

a. Calculate the induced voltage between the axis and rim of the disk. Express your answer in volts.

b. Draw a diagram indicating the directions of field, rotation, and induced e.m.f.

4. A copper disk 10 cm. in radius is placed inside a long solenoid of radius 11 cm. with its axis coincident with that of the solenoid. The solenoid is 1 meter long, has 1,000 turns, and carries a steady current of 2 amp. The copper disk is rotated about its axis with a uniform angular velocity of 1,200 r.p.m. Calculate the e.m.f. induced between slip rings connected to the rim and axis in volts.

5. The flux linking a 100-turn coil connected in series with a ballistic galvanometer is changed suddenly and the maximum angular deflection of the galvanometer is 0.01 radian. The sensitivity of the galvanometer is

10 radians/coulomb and the resistance of the circuit is 10 ohms. Calculate the change in flux through the coil.

6. A large circular coil of N turns and radius b carries a current i and is rotated uniformly about a horizontal diameter. At the center of this coil is a small fixed horizontal circular turn of radius a. Calculate the e.m.f. induced in the small loop as a function of the time. What is the angle between the planes of the coils when the e.m.f. is a maximum?

7. Given N turns of wire connected in series. Under what conditions would you expect the inductance of such a coil to be proportional to the number of turns? How would you arrange these turns to obtain the maximum inductance?

8. A solenoid 3.0 cm. in radius and 90 cm. long is wound closely and uniformly with 20 turns per centimeter of length. What is the self-inductance of this solenoid in henrys?

9. A closely wound circular coil of 60 turns is wound around the central portion of the solenoid of Prob. 8. Assuming that all the flux produced by the solenoid links this coil, calculate the mutual inductance of the system. Would this be altered if the 60-turn coil were short-circuited by connecting its ends together?

10. In Prob. 9 suppose the solenoid carries a current given by

$$i = 2 \sin 2\pi\nu t \qquad \text{(amperes)}$$

where $\nu = 60$ cycles/sec. and t is the time in seconds. Compute the voltage induced in the 60-turn coil as a function of time when this coil is on open circuit.

11. Given two concentric coplanar circular coils A and B of N_1 and N_2 turns, respectively. Let r_1 be the radius of coil A and r_2 that of coil B, and suppose $r_1 >> r_2$, so that one may assume the magnetic field produced by coil A to be uniform over the area of coil B. Deduce an expression for the mutual inductance of the system.

12. Solve Prob. 11 for the case in which the two coils are coaxial, their planes being separated by a distance R which is much larger than r_1.

13. Two long straight parallel wires forming part of a circuit are separated by a distance d and carry equal steady currents in opposite directions. Neglecting the flux inside the wires, calculate an expression for the inductance per unit length of this part of the circuit.

14. A coil carrying a steady current i is moved from a position in which no external magnetic field exists to a position where an external flux Φ links the coil. During this process the current is maintained constant with the help of suitable batteries. Show that the work done by these batteries against induced e.m.fs. is $i\Phi$.

15. Starting from the expression for the energy density in a magnetic field, compute the magnetic energy stored per unit length *inside* a circular metallic wire of radius r when it carries a steady current i. From this derive an expression for the contribution to the inductance of a circuit (per unit length) of the flux inside the conductor.

16. Proceeding according to the scheme of Prob. 15, derive an exact formula for the inductance per unit length of the coaxial cable of Prob. 31,

Chap. V. Compare your result with that given by Eq. (16), and state the conditions under which Eq. (16) is a good approximation.

17. A transmission line consists of two parallel wires each of circular section, radius a, separated by a distance b between the axes of the wires. Deduce an expression for the inductance per unit length of the transmission line. Assume $b >> a$.

18. Prove that, if two coils are connected in series, the inductance of the system is given by $L_1 + L_2 \pm 2M$, where L_1 and L_2 are the self-inductances of the two coils and M the mutual inductance of the system. Under what conditions is the $+$ sign valid?

19. A rectangular coil of width a and length l having N turns is placed between two very long parallel conductors of separation $3a$ forming part of a circuit. The coil is centered between the conductors, the long sides l being parallel to the conductors and the plane of the coil lying in the plane of the parallel conductors.

Derive an expression for the mutual inductance of the system.

20. Two fixed coils of self-inductances L_1 and L_2 and mutual inductance M carry steady currents i_1 and i_2, respectively. Prove that the energy stored in the magnetic field is given by

$$U_m = \tfrac{1}{2}L_1i_1^2 + \tfrac{1}{2}L_2i_2^2 \pm Mi_1i_2$$

the sign of the last term depending on the relative directions of the currents.

21. Using the fact that the expression for the magnetic field energy as given in Prob. 20 must be independent of the manner in which the currents i_1 and i_2 are established, prove that the proportionality constants K of Eqs. (17) and (18) are equal.

CHAPTER VII

ELEMENTARY ALTERNATING-CURRENT CIRCUITS

In this chapter we shall investigate the behavior of the simplest circuits in which there are e.m.fs. and currents varying with the time, and particularly we shall consider the case of sinusoidal time variations. We have already seen that one can generate a sinusoidally varying e.m.f. in a coil by rotating it with constant angular velocity in a uniform magnetic field [Eq. (9), Chap. VI], and we shall study the steady-state behavior of simple circuits connected to the terminals of such an a.c. generator. We shall assume that the generators employed are sufficiently large so that we may take the terminal voltage as independent of the current drawn from them.

The whole theory of a.c. circuits is based on the application of the Faraday induction law, which provides the extension of the Kirchhoff rules employed in the case of steady currents. In the latter case the statement that the sum of the voltage drops around any closed surface equals zero is equivalent to the fundamental law of electrostatics

$$\oint \mathcal{E}_s \, ds = 0$$

and the evaluation of the integral in terms of iR drops and impressed e.m.fs. leads to the d.c. circuit equations. For the case of currents varying with the time, this equation takes the more general form given by the Faraday law, *viz.*,

$$\oint \mathcal{E}_s \, ds = -\frac{d}{dt}(N\Phi) = -L\frac{di}{dt}$$

where the evaluation of the integral leads to the sum of the instantaneous voltage drops around the circuit. Thus we shall be led to circuit equations which are entirely similar to those for steady currents, differing only in the presence of the $L(di/dt)$ term. There are two common modes of interpretation of this term: (1) One puts it on the left-hand side of the equation and

treats it as an additional voltage drop in the circuit. In this manner one retains the original form of the Kirchhoff law for the a.c. case. (2) One can look upon this term as representing a "back" e.m.f. which must be subtracted from the impressed e.m.fs. to obtain the net e.m.f. acting around the circuit. Both procedures lead, of course, to identical equations.

The fundamental assumption which is made in applying the above reasoning to actual circuits is *that the current variations are sufficiently slow so that, at a given instant of time, the electric and magnetic fields are essentially the same as would be produced by the corresponding steady currents and charges.* Thus, for example, we consider the current to be the same at every point in a series circuit at a given instant of time, so that the idea of inductance may be employed to describe the rate of change of magnetic flux through the circuit. An exact criterion for the validity of the above assumption must be delayed to a later chapter, and the phenomena for which this assumption is valid are called *quasi-stationary.*

36. The Simplest Alternating-current Circuits.—First let us consider the simplest possible case, a noninductive resistance R connected across the terminals of a generator whose terminal voltage is given by $E = E_0 \sin \omega t$, where $\omega = 2\pi n$, n being the frequency of this impressed sinusoidal e.m.f. Assuming the validity of Ohm's law for varying currents, we have for the sum of the voltage drops around the series circuit,

$$-E + iR = 0$$

or

$$E = iR \tag{1}$$

In this equation, which is identical in form with that employed to describe a similar d.c. circuit, both E and i vary with the time, so that Eq. (1) gives the relation between *instantaneous* current and *instantaneous* generator e.m.f. From Eq. (1) we find for the current i, as a function of the time,

$$i = \frac{E}{R} = \frac{E_0}{R} \sin 2\pi nt = I \sin 2\pi nt \tag{2}$$

where $I = E_0/R$ is the maximum value of the current. Equation (2) shows that a sinusoidally varying current flows which is in phase with the sinusoidally varying voltage drop across the resistance.

Now let us consider the case of a coil of inductance L and negligible resistance connected to the generator terminals (Fig. 84). At any instant of time there is an e.m.f. induced in the coil equal to $-L\frac{di}{dt}$, so that Kirchhoff's laws applied to this circuit give

$$E - L\frac{di}{dt} = 0$$

or

$$E = L\frac{di}{dt} \tag{3}$$

The rate of change of current with time at any instant is propor-

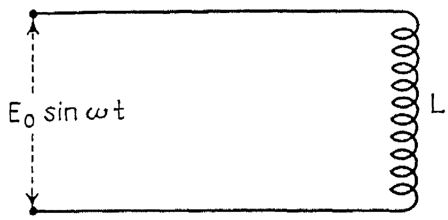

FIG. 84.

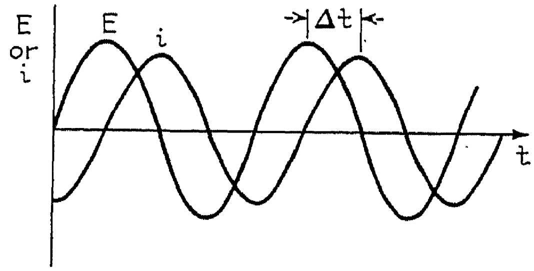

FIG. 85.

tional to the instantaneous voltage applied to the coil. If we write Eq. (3) in the form

$$\frac{di}{dt} = \frac{E}{L} = \frac{E_0}{L}\sin \omega t$$

a direct integration yields for the current

$$i = -\frac{E_0}{\omega L}\cos \omega t = -\frac{E_0}{2\pi n L}\cos 2\pi n t \tag{4}$$

The current varies as $(-\cos \omega t) = \sin\left(\omega t - \frac{\pi}{2}\right)$, so that again we have an alternating current of the same frequency as the generator voltage, but in this case the current is not in phase with the voltage, lagging behind it by $90° = \pi/2$ radians. In Fig. 85 are plotted the voltage drop E across the inductance and the current i flowing in it as functions of time t. From the plot one sees that the current reaches its maximum value later than the voltage. The *time lag* Δt is also shown. This time lag Δt can be obtained as follows: We write the current as

$$i = \frac{E_0}{\omega L} \sin\left(2\pi nt - \frac{\pi}{2}\right) = \frac{E_0}{\omega L} \sin\left[2\pi n\left(t - \frac{1}{4n}\right)\right]$$

so that $\Delta t = 1/4n$. The quantity $\omega L = 2\pi nL$ is known as the inductance *reactance* of the coil at the frequency n and is usually denoted by X_L. We can write for the maximum current

$$I_{\max} = \frac{E_0}{\omega L} = \frac{E_0}{X_L} \tag{5}$$

As a final example, consider the case of a condenser of capacitance C connected to the terminals of the generator (Fig. 86). The charge on the condenser plates will vary with time, and we can think of the potential difference between the plates varying with the time. The flow of charge to and away from the condenser plates constitutes an electric current in the wires connected to the condenser given by $i = dq/dt$. We now assume that, as the charge q on the plates varies, the potential drop across the condenser is given at any instant of time by the ratio of the charge on the plates at that instant of time to the capacity of the condenser. Thus the static relation $V = q/C$ is assumed valid even when q varies with the time. This will be true as long as the time necessary for a charge to distribute itself uniformly over the plates is very small compared with optical frequencies.

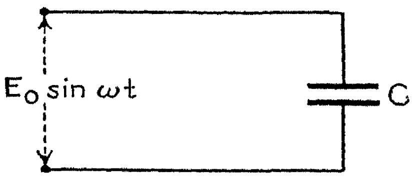

FIG. 86.

According to the above, we write for the charge on the condenser at any time as

$$q = CE_0 \sin \omega t \tag{6}$$

and hence the current is given by

$$i = \omega CE_0 \cos \omega t \tag{7}$$

The current varies as $\cos \omega t = \sin\left(\omega t + \frac{\pi}{2}\right)$; consequently we have a sinusoidally alternating current of the same frequency as the generator voltage but not in phase with it, the current leading the voltage by $90° = \pi/2$ radians.

In Fig. 87 are plotted the voltage drop E across the condenser and the current i in the circuit as functions of the time t. One sees that the current reaches its maximum value before the

voltage becomes maximum; the *time lead* Δt is shown and equals $1/4n$. The quantity $1/\omega C = 1/2\pi nC$ is known as the *capacitive reactance* of the condenser and is denoted by X_C. The maximum current can be written in the form

$$I_{\max} = \omega C E_0 = \frac{E_0}{X_C} \tag{8}$$

Summarizing, we have the following important results:

1. When an alternating current flows through a noninductive resistance, the voltage drop across the resistance is given by iR, and the current is in phase with this voltage drop.

2. The voltage drop across an inductance carrying an alternating current i is given by $i\omega L = iX_L$, and the current *lags* the voltage drop by 90°.

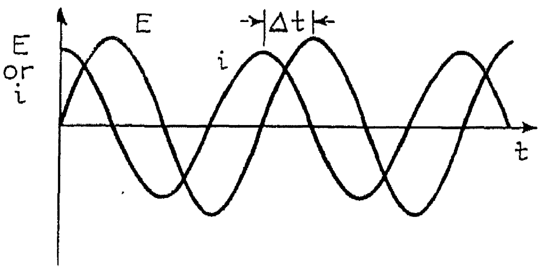

FIG. 87.

3. When a condenser is connected in an a.c. circuit, the voltage drop across the condenser is given by $i/\omega C = iX_C$, and the current *leads* the voltage drop by 90°.

37. Vector Representation of Sinusoidal Functions.—In our following discussion we shall have occasion to add two or more sinusoidally varying voltages when we apply Kirchhoff's laws to a series a.c. circuit. Hence we digress for a moment to study the laws of addition. For simplicity consider the sum of two alternating voltages of the same frequency but different phases

$$\begin{aligned} E_1 &= A \sin \omega t \\ E_2 &= B \sin (\omega t - \end{aligned} \tag{9}$$

Here A and B are the maximum values of E_1 and E_2, respectively, and δ is the phase difference between the voltages. The sum of these terms E will be a sine function of the same frequency but not in phase with either E_1 or E_2. Let this sum be

$$E = E_1 + E_2 = C \sin (\omega t - \epsilon) \tag{10}$$

and our problem is to determine the amplitude and phase C and ϵ in terms of A, B, and δ.

We have

$$E = E_1 + E_2 = A \sin \omega t + B \sin (\omega t - \delta) = A \sin \omega t + B \cos \delta \sin \omega t - B \sin \delta \cos \omega t \tag{11}$$

where we have used the relation

$$\sin(\alpha - \beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta$$

Equation (11) can be written as

$$E = (A + B\cos\delta)\sin\omega t - B\sin\delta\cos\omega t \tag{12}$$

and we wish to identify this with Eq. (10). Equation (10) can be written as

$$E = C\sin(\omega t - \epsilon) = C\cos\epsilon\sin\omega t - C\sin\epsilon\cos\omega t \tag{13}$$

Comparing Eqs. (12) and (13), we must have

$$\left.\begin{aligned} C\sin\epsilon &= B\sin\delta \\ C\cos\epsilon &= A + B\cos\delta \end{aligned}\right\} \tag{14}$$

from which we can calculate C and ϵ. Squaring these equations and adding, there follows

$$C = \sqrt{(A + B\cos\delta)^2 + B^2\sin^2\delta} = \qquad \overline{\cos\delta} \tag{15}$$

and dividing the first by the second

$$\tan\epsilon = \frac{B\sin\delta}{A + B\cos\delta} \tag{16}$$

Equations (15) and (16) are *exactly* the formulas for the sum of two vectors A and B which make an angle δ with each other (Fig. 88). Thus we have the fundamental theorem: *The sum of two sinusoidal functions of the same frequency is a sinusoidal function of the same frequency, and the amplitude and phase of the sum can be obtained from those of the two given functions by vector addition.* We can therefore represent sine functions by vectors, the lengths of the vectors being the amplitudes of the sine functions and the angles between the vectors being the phase differences between the corresponding sine functions. One can readily show that the addition of more than two functions is also equivalent to adding vectors.

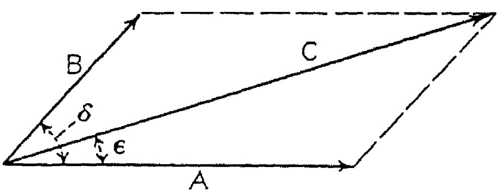

FIG. 88.

38. The Simple Series Circuit.—Consider a simple series circuit consisting of resistance R, inductance L, and capacitance

C, connected to the terminals of an a.c. generator as shown in Fig. 89. Applying Kirchhoff's laws to this circuit (the sum of the voltage drops around the circuit must be zero), we have

$$-E_0 \sin \omega t + iR + L\frac{di}{dt} + \frac{q}{C} = 0 \tag{17}$$

or

$$L\frac{di}{dt} + iR + \frac{q}{C} = E_0 \sin \omega t \tag{18}$$

and the solution of this equation yields the current i in the circuit and the charge q on the condenser. The complete solution of this equation yields two terms; one is a transient current which depends on the initial conditions and soon dies out and the other is a steady-state current which persists as long as the applied voltage acts in the circuit. We shall confine ourselves to the *steady state* in this section, and, since the steady-state current has the same frequency as the applied voltage, we can utilize the results of the previous sections to obtain the solution.

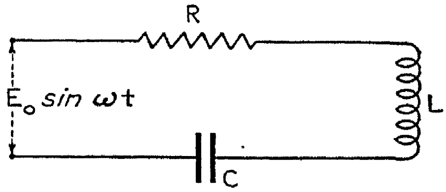

FIG. 89.

Equation (18) states that the applied voltage ($E_0 \sin \omega t$) is the sum of the voltage drops across inductance, resistance, and condenser, and we shall add these with the help of a vector representation. We can write symbolically

$$\vec{E} = \vec{V}_L + \vec{V}_R + \vec{V}_C$$

where the $\vec{V}$'s are the vectors representing the drops across inductance, resistance, and capacitance. From the results of Sec. 36 we have for the magnitudes of these vectors:

$$V_L = \omega LI$$
$$V_R = IR$$
$$V_C = \frac{I}{\omega C}$$

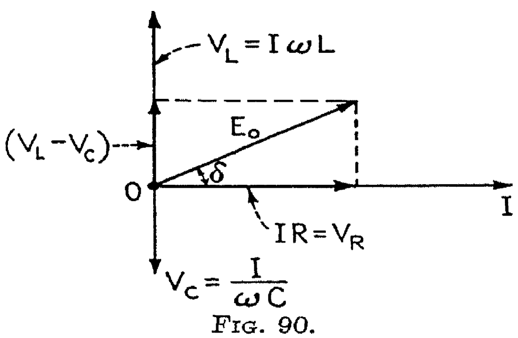

FIG. 90.

where I is the amplitude of the current i. Figure 90 shows the *vector diagram* for the circuit. The vector I represents the common current in the circuit, the

vector $IR = V_R$ is in phase with I (zero angle between them), the vector $\omega LI = V_L$ is 90° ahead of I (considering counterclockwise rotation positive, as usual), and $I/\omega C = V_C$ is 90° behind I, according to the results previously found. Evidently we have

$$\left.\begin{aligned} E_0 &= I\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2} = I\sqrt{R^2 + (X_L - X_C)^2} \\ &\text{and} \\ \tan\delta &= \frac{\omega L - \dfrac{1}{\omega C}}{R} = \frac{X_L - X_C}{R} \end{aligned}\right\} \tag{19}$$

The first of these equations relates the amplitude of the current I to that of the applied voltage E_0. The second equation gives the *phase angle* δ between current and applied e.m.f. This is a positive angle when the current lags the voltage and is negative when the current leads the voltage. The quantity

$$\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}$$

is called the *impedance* of the circuit and is denoted by the letter Z.

From Eqs. (19) we see that the instantaneous current i is given by

$$i = \frac{E_0}{Z}\sin(\omega t - \delta) \tag{20}$$

if $E = E_0 \sin \omega t$. Here the impedance Z is

$$Z = \sqrt{R^2 + (X_L - X_C)^2} = \sqrt{R^2 + \left(2\pi nL - \frac{1}{2\pi nC}\right)^2} \tag{21}$$

and δ is given by the second of Eqs. (19). *Equation* (20) *is the steady-state solution of Eq.* (18). Note that impedance and reactance are measured in the same units as resistance. From the expression for $\tan\delta$ we see that the current will be in phase with the applied voltage if $X_L = X_C$. When this condition exists, the circuit is said to be in *resonance*, and the current has the maximum value possible (Z is a minimum). At resonance, then, Eq. (20) becomes

$$i = \frac{E_0}{R}\sin\omega t \qquad \text{(at resonance)} \tag{22}$$

when $\omega L = 1/\omega C$. This resonance condition written in terms of frequency becomes

$$n = \frac{1}{2\pi}\sqrt{\frac{1}{LC}} \tag{23}$$

Further discussion of the series resonant circuit is left to the problems.

39. Energy Considerations for the Series Circuit.—Now let us examine the energy relations which must exist in the series circuit which we have been examining. First, the power input to the circuit from the generator Ei is not constant but varies with the time. There will be, however, an average rate of doing work which we must calculate. In the resistance R the power consumed (i^2R) will vary with the time so that the rate of heating is not constant. Again there is an average value which must be computed. In the inductance L there will be an instantaneous power consumption $Li(di/dt)$ which represents the rate at which the energy in the magnetic field increases or decreases. This averages to zero over a cycle as we shall see in a moment. Finally, there will be an instantaneous power consumption by the condenser equal to $(q/C)i$ representing the rate at which the energy in the electric field between the plates is increasing or decreasing. As in the case of the magnetic energy, this averages to zero over a cycle.

In the resistance the rate of heating at any instant is

$$i^2R = I^2R\sin^2(\omega t - \delta) = I^2R\sin^2\theta \tag{24}$$

utilizing Eq. (20) with $I = E_0/Z$ and $\theta = \omega t - \delta$.
The average rate of heating (averaged over one period of oscillation) is accordingly

$$\overline{i^2R} = I^2R\left(\frac{1}{2\pi}\int_0^{2\pi}\sin^2\theta\,d\theta\right) \tag{25}$$

since, in one period, θ increases by 2π and by definition the average value of any function $f(x)$, let us say, over an interval 0 to a is

$$\overline{f(x)} = \frac{1}{a}\int_0^a f(x)\,dx$$

The integral in Eq. (25) may be evaluated simply as follows: From the similarity of the sine and cosine functions we must have

$$\int_0^{2\pi} \sin^2 \theta \, d\theta = \int_0^{2\pi} \cos^2 \theta \, d\theta = \int_0^{2\pi} (1 - \sin^2 \theta) \, d\theta = 2\pi - \int_0^{2\pi} \sin^2 \theta \, d\theta$$

so that

$$\int_0^{2\pi} \sin^2 \theta \, d\theta = \pi$$

Substituting this value in Eq. (25) there follows for the average rate of heating

$$P_R = \overline{i^2 R} = \frac{I^2}{2} R \tag{26}$$

or one-half the maximum rate. From this relation we see that the average value of the square of the current i is $I^2/2$. It is convenient to introduce the idea of the square root of this average square of the current. This quantity $\sqrt{\overline{i^2}}$ is called the *root-mean-square* (r.m.s.) value of the current or the *effective* current, and this is the value which would be read on an a.c. ammeter. Thus we can write

$$i_{\text{eff}} = \sqrt{\overline{i^2}} = \frac{1}{\sqrt{2}} I = 0.707I \tag{27}$$

The effective value of a sinusoidally varying current is $1/\sqrt{2}$ times the maximum value of the current. An analogous relation holds for the effective value E_{eff} of a sinusoidally varying voltage. In terms of effective current, the average rate of heating in the resistance is simply $i_{\text{eff}}^2 R$.

The rate of increase of energy in the magnetic field of the inductance is given at any instant of time by

$$Li\frac{di}{dt} = \omega L I^2 \sin (\omega t - \delta) \cos (\omega t - \delta) = \omega L \frac{I^2}{2} \sin 2(\omega t - \delta)$$

In terms of the effective current i_{eff}, this becomes simply

$$P_L = \omega L i_{\text{eff}}^2 \sin 2(\omega t - \delta) = i_{\text{eff}}^2 X_L \sin 2(\omega t - \delta) \tag{28}$$

The average power input to the inductance is thus zero since the average value of the $\sin 2(\omega t - \delta) = 0$. Energy is alternately stored up in the magnetic field and returned to the generator.

A similar state of affairs exists in the condenser. The power input to the condenser is

$$P_C = \frac{q}{C}i = -\frac{I^2}{\omega C}\sin(\omega t - \delta)\cos(\omega t - \delta) = -\frac{I^2}{2\omega C}\sin 2(\omega t - \delta)$$

or

$$P_C = -\frac{i_{\text{eff}}^2}{\omega C}\sin 2(\omega t - \delta) = -i_{\text{eff}}^2 X_C \sin 2(\omega t - \delta) \tag{29}$$

Energy is alternately stored up in the condenser while it is charging and returned to the generator during discharge. The average power consumption is zero.

The power input to the whole circuit from the generator is at any instant

$$P = Ei = E_0 I \sin \omega t \sin(\omega t - \delta) = E_0 I \cos\delta \sin^2 \omega t - E_0 I \sin\delta \sin\omega t \cos\omega t \tag{30}$$

The average power input is therefore

$$\bar{P} = \frac{E_0 I}{2}\cos\delta = E_{\text{eff}} i_{\text{eff}} \cos\delta \tag{31}$$

since the last term averages to zero.

The ratio of the average power input to a system to the product of the effective voltage across the system and the effective current flowing is called the power factor of the system. For the case of sinusoidal alternating currents Eq. (31) shows us that

$$\text{p.f.} = \cos\delta \tag{32}$$

The relations just derived have a simple interpretation based on a vector diagram, as shown in Fig. 91. Here we show the effective voltage and current instead of maximum values and the phase difference δ. The average power input, as given by Eq. (31), is clearly the product of the r.m.s. voltage and the component of current in phase with this voltage. The second term in Eq. (30) for the instantaneous power can be put in the form $(E_{\text{eff}} i_{\text{eff}} \sin\delta)\sin 2\omega t$, and this quantity is often referred to as the "wattless" or "reactive" power. For certain values of t, Eq. (30) can yield negative values of the power input and at these times energy is transferred from the circuit to the seat of e.m.f.

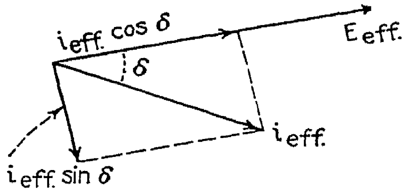

FIG. 91.

40. Free Oscillations of an L C Circuit; Simple Transients.—There is a simple important case in which steady-state alternating currents may be set up without the aid of an external seat of e.m.f. Consider the circuit of Fig. 92, consisting of a condenser and inductance in series, and let us suppose that the resistance is negligible. We charge the condenser to an initial difference of potential V_0, so that it has an initial charge q_0, and close the switch S. Kirchhoff's laws give

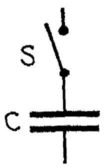

FIG. 92.

$$L\frac{di}{dt} + \frac{q}{C} = 0$$

or, since $i = dq/dt$,

$$\frac{d^2q}{dt^2} + \frac{q}{LC} = 0 \tag{33}$$

This is the equation of simple harmonic motion, and the solution is

$$q = q_0 \cos \omega t \tag{34}$$

and

$$i = \frac{dq}{dt} = -\omega q_0 \sin \omega t = -\omega C V_0 \sin \omega t \tag{35}$$

since at $t = 0$, $q = q_0$ and $i = 0$. $\omega = 1/\sqrt{LC}$, so that we obtain free oscillations of the system with a frequency n given by

$$n = \frac{1}{2\pi}\sqrt{\frac{1}{LC}} \tag{36}$$

While the system is oscillating, there is a constant interchange of energy between the electric field of the condenser and the magnetic field of the inductance, the total energy remaining constant. This is analogous to the case of mechanical oscillations with the interchange of potential and kinetic energy. If resistance is present, the oscillations die out as they do when friction is present in the mechanical case.

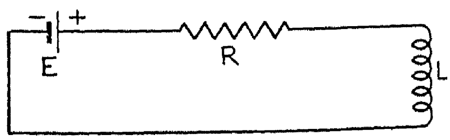

FIG. 93.

Thus far we have considered only steady-state solutions of the circuit equations, and we now turn to two simple cases of nonsteady or transient currents. Consider first the circuit shown in Fig. 93. At any instant of time we have

$$E - L\frac{di}{dt} - iR = 0$$

or

$$L\frac{di}{dt} + Ri = E \tag{37}$$

To integrate this equation, it is convenient to make the substitution

$$i = i_1 + \frac{E}{R}$$

Substituting in Eq. (37), we find

$$L\frac{di_1}{dt} + Ri_1 = 0 \tag{38}$$

so that the constant term E has been eliminated. Equation (38) can now be directly integrated, yielding

$$i_1 = c\, e^{-\frac{R}{L}t}$$

where c is an arbitrary constant and the complete solution becomes

$$i = \frac{E}{R} + c\, e^{-\frac{R}{L}t} \tag{39}$$

Equation (39) shows that the current is the sum of a steady-state term E/R (Ohm's law) and a transient term $c\, e^{-\frac{R}{L}t}$ which depends on initial conditions. For example, let us consider the case in which a switch is closed to establish the closed circuit at $t = 0$. We then have $i = 0$ when $t = 0$, and Eq. (39) yields for the integration constant c

$$c = -\frac{E}{R}$$

so for that case

$$i = \frac{E}{R}(1 - e^{-\frac{R}{L}t}) \tag{39a}$$

A plot of current versus time is shown in Fig. 94. Now let us suppose that a steady current flows in the circuit of Fig. 93 and that at $t = 0$ the e.m.f. is removed. At all subsequent times there is no external e.m.f. acting in the circuit; consequently we place $E = 0$ in Eq. (37), obtaining

$$L\frac{di}{dt} + Ri = 0$$

the solution of which is

$$i = i_0 e^{-\frac{R}{L}t} \tag{39b}$$

where i_0 is the value of the current when the e.m.f. is removed.

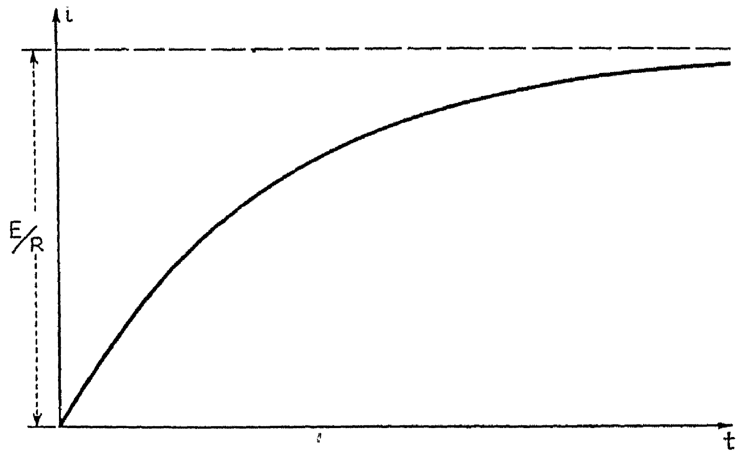

FIG. 94.

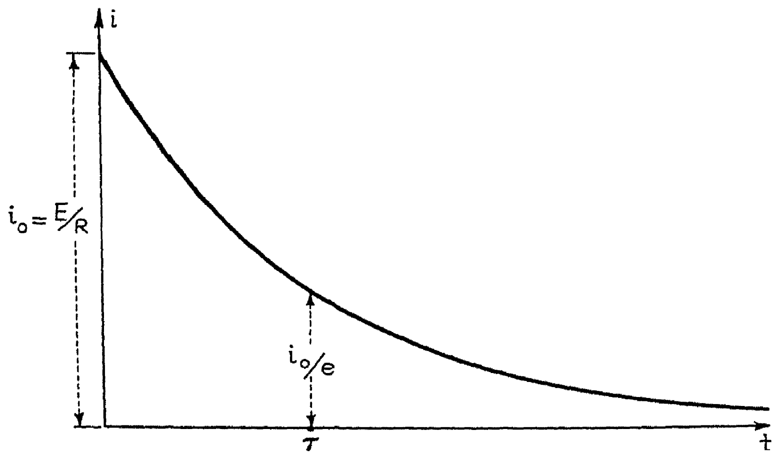

FIG. 95.

A plot of i versus t is shown in Fig. 95. The time elapsed before the current in the circuit drops to $1/e$th of its initial value is called the *time constant* of the circuit. Denoting this by τ we have

$$\tau = \frac{L}{R} \tag{39c}$$

As a second and final example of transient currents let us consider the discharge of a condenser through a resistance (Fig. 96). Suppose that initially the condenser carries a charge q_0, that the potential difference $V_0 = q_0/C$, and that at $t = 0$ a switch is closed establishing the circuit shown. The discharge is given by the equation (Kirchhoff's rules)

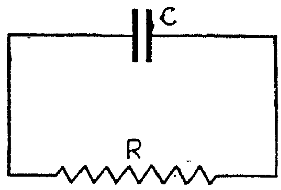

FIG. 96.

$$\frac{q}{C} + iR = 0$$

or, since $i = dq/dt$,

$$\frac{dq}{dt} + \frac{q}{RC} = 0 \tag{40}$$

Integrating this equation, we find

$$q = q_0 e^{-\frac{t}{RC}} \tag{41}$$

and from this the current

$$i = \frac{dq}{dt} = -\frac{q_0}{RC} e^{-\frac{t}{RC}} = -\frac{V_0}{R} e^{-\frac{t}{RC}} \tag{42}$$

Both current and charge decay exponentially with the time, the curve being similar to that of Fig. 95. The time constant of such a circuit is

$$\tau = RC \tag{43}$$

Problems

(*All voltages and currents are r.m.s. values unless otherwise stated.*)

1. A coil of unknown inductance and of resistance 37.7 ohms is connected to a 230-volt, 60 cycle, a.c. line, and a current of 4.7 amp. flows through the coil.

a. Calculate the reactance and inductance of the coil.

b. Calculate the phase difference between the voltage across the coil and the current through it.

c. What current would flow if the frequency of the applied voltage were 30 cycles per second?

2. A coil takes a current of 25 amp. when connected to 220-volt, 60-cycle mains. If this same coil in series with a 5-ohm resistance is connected to a 110-volt d.c. line, the current is 17 amp. Compute the resistance and inductance of the coil.

3. A resistance draws 10 amp. when connected to a 110-volt, 60-cycle line. How big a condenser must be connected in series with the resistance so that the current drop to 5 amp.? What is the voltage across the condenser and across the resistance for this connection?

4. A series circuit consists of a 300-ohm resistance, an inductance whose reactance is 400 ohms at 60 cycles, and a condenser of 500 ohms reactance at 60 cycles. A 60-cycle voltage of 500 volts is applied to the circuit.

a. Calculate the inductance in henrys and the capacity of the condenser in microfarads.

b. What is the impedance of the circuit?

c. What steady-state current flows in the circuit?

d. What is the phase angle between the current and the applied voltage? Does the current lead or lag the voltage?

e. What size condenser would be needed to produce resonance? What current would flow under these conditions?

5. A series circuit consisting of a resistance R, inductance L, and capacitance C, is connected to a generator of variable frequency so that the applied voltage is given by $E = E_0 \sin 2\pi nt$, where E_0 is constant at all frequencies.

a. Calculate expressions for the impedance of the circuit as a function of frequency n if $L = 0.1$ henry and $C = 1\ \mu$f for the following values of resistance: $R = 100$ ohms; $R = 10$ ohms; $R = 1$ ohm.

b. Plot the current as a function of frequency for the three values of resistance.

c. What is the frequency at which resonance occurs? What current flows at resonance for $R = 1$ ohm, $E_0 = 10\sqrt{2}$ volts? For what values of frequency is the current one-half its maximum value?

6. A coil of resistance 2 ohms and inductance 0.1 henry is used in series with a condenser to show resonance. The only source of e.m.f. available is a 110-volt, 60-cycle line. What is the necessary capacity of the condenser? If the condenser is designed to stand an effective voltage of 500 volts, what resistance must be inserted in the circuit to limit the drop across the condenser to this value?

7. An alternating voltage of constant amplitude and variable frequency is applied to a series R, L, C circuit. Deduce an expression for the frequency at which the voltage drop across the condenser is a maximum. Is this higher or lower than the resonant frequency of the circuit?

8. For the circuit of Prob. 7, deduce an expression for the frequency at which the voltage drop across the inductance is a maximum.

9. The equations of two alternating e.m.f.s are $E_1 = 150 \sin 377t$ and $E_2 = 150 \sin (377t + 60°)$. If these e.m.fs. are in series, what is the equation of their resultant? What is the phase angle between the resultant and each of the two e.m.fs.? At the instant of time when the resultant is zero, what are the values of E_1 and E_2?

10. An alternating current of amplitude 0.10 amp. is rectified so that current flows only during the positive half cycles (half-wave rectification). Calculate the average and r.m.s. values of the current.

11. Two coils have resistances of 10 and 16 ohms and inductances of 0.02 and 0.4 henry, respectively. If connected in series across 220-volt, 60-cycle mains, what current will they take? What is the power factor of the circuit? What is the phase difference between the voltage drops across the two coils?

12. A coil of 2.7 ohms resistance and variable inductance is connected in series with a noninductive resistance across a 220-volt, 60-cycle line. The circuit is so adjusted that the drop across the coil is 150 volts and the power it absorbs is 250 watts.

What is the value of the noninductive resistance?

13. A noninductive resistance of 25 ohms in series with a condenser absorbs 968 watts when connected to a 220-volt, 60-cycle line. What current will this circuit take when connected to a 110-volt, 25-cycle line? What power will it absorb?

14. A coil is connected in series with a condenser across 220-volt, 60-cycle mains. The circuit absorbs 650 watts at a power factor of 0.87 and is so adjusted that the drops across coil and condenser are equal. What are these voltages?

15. A series circuit with $R = 10$ ohms, $L = 0.1$ henry, and $C = 30$ μf is connected to 110-volt, 60-cycle mains. At the instant at which the impressed e.m.f. is zero, what is the energy in the condenser? When there is no energy in the condenser, what is the energy in the magnetic field of the inductance?

16. A coil takes 250 watts at a power factor of 0.1 when connected to a 220-volt, 60-cycle line. What capacity must be connected in series with this coil so that it takes the same power from a 110-volt, 60-cycle line? What is the power factor of the latter circuit?

17. A coil has a resistance of 1 ohm and an inductance of 0.1 henry.

a. What is the time constant of the coil?

b. Plot the values of the current during the first second after an e.m.f. of 10 volts is impressed on the coil.

c. Plot the rate at which energy is supplied to the coil during this second.

d. What is the energy of the magnetic field 0.1 sec. after the switch is closed?

18. A coil has a resistance of 5 ohms and an inductance of 0.15 henry. When carrying a current of 25 amp. the impressed e.m.f. is suddenly replaced by a noninductive resistance of 10 ohms.

a. What per cent of the initial magnetic energy is ultimately dissipated in the 10-ohm resistance?

b. What is the initial voltage across the 10-ohm resistance?

19. A 10-μf condenser which has been charged to a potential of 200 volts is discharged by connecting a 1,000-ohm resistance across its terminals.

a. What is the initial energy stored in the condenser and the initial value of the discharge current?

b. What is the current when the charge on the condenser has fallen to one-half of its initial value?

c. Compute the total heat generated in the resistance by evaluating $\int_0^\infty i^2R\,dt$ and compare with the answer to part *a*.

20. A condenser of capacity C is charged through a resistance R by connecting the two in series to the terminals of a battery of e.m.f. E.

Derive formulas for the charge on the condenser and for the charging current as functions of the time.

CHAPTER VIII

DISPLACEMENT CURRENT AND ELECTROMAGNETIC WAVES

In Chap. V we formulated the laws of the stationary magnetic field produced by steady electric currents, and now we must extend these laws to cover nonsteady or transient phenomena. The concept of inductance and the circuit laws discussed in the last chapter have been based on the assumption that the Ampère circuital law for steady currents remained valid for currents varying with the time. We are now ready to investigate the range of validity of this assumption. From the fundamental law of conservation of electric charge, as expressed in the equation of continuity, we have seen that the lines of current flow always close on themselves for the *steady state;* consequently one has to deal only with *closed* circuits. For nonsteady, transient currents this is no longer the case, and one may have current flowing in an *open circuit* as, for example, in the case of a condenser being charged or discharged. Our first task is to formulate the equation of continuity for nonsteady currents.

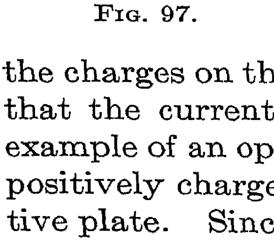

FIG. 97.

41. The Equation of Continuity for Charge and Current.—Let us first examine the simple case of a condenser which is being discharged (Fig. 97). Suppose that at some instant of time the charges on the condenser plates are $+q$ and $-q$, as shown, and that the current flowing at this instant is i. This is a typical example of an open circuit, the lines of current flow starting at the positively charged condenser plate and terminating on the negative plate. Since, by definition, the current i is the rate at which charge crosses a cross section of the conducting wire normal to the direction of flow, we see that the law of conservation of charge requires that the rate of increase of charge on a condenser plate just equals the current flowing into that plate. We have already employed this relation between the charge on a condenser and the current in our study of a.c. circuits.

To formulate this continuity law mathematically, let us imagine that we construct a *closed* surface which completely encloses one of the condenser plates in Fig. 97, let us say the positive plate. The current flowing *outward* across this surface is i, which equals the rate of decrease of charge on the plate, and we may write

$$i = -\frac{dq}{dt} \tag{1}$$

In Eq. (1) the left-hand side refers to the current flowing *out* of the volume enclosed by the surface. We can now formulate Eq. (1) more generally. The total current flowing out of a closed volume can be written as $\int j_n\, dS$, where j_n is the normal component of the current density at a point where the element of area dS is located and the integral extends over the whole closed surface (compare Eqs. (3) and (6), Chap. IV). Equation (1) can then be written in the form

$$\int_{\substack{\text{closed}\\ \text{surface}}} j_n\, dS = -\frac{dq}{dt} \tag{2}$$

This is the general form for the equation of continuity for stationary bodies in which q represents the charge inside the closed surface. Equation (2) reduces to Eq. (6) of Chap. IV for the steady state, since in this case the right-hand side vanishes. If the charge inside the volume enclosed by the surface is distributed with a space density ρ, we can write Eq. (2) in the form

$$\int_{\substack{\text{closed}\\ \text{surface}}} j_n\, dS = -\int \frac{\partial \rho}{\partial t}\, dv \tag{3}$$

where the integral on the right-hand side extends throughout the volume. We have used the partial derivative $\partial\rho/\partial t$ to indicate the rate of change of charge density at a fixed point (where dv is located) inside the volume.

42. The Maxwell Displacement Current.—If we now inquire into the question of the magnetic field produced by transient or nonsteady currents, we find that the Ampère circuital law, which states that the magnetomotive force around a closed path is equal to 4π times the current flowing across any surface

of which the closed path is a boundary, breaks down for open circuits. We can carry through the discussion best for the simple case of a condenser being charged. In Fig. 98 are shown the condenser and a closed circular path ↺ surrounding the wire leading to the positive plate. The magnetomotive force $\oint H_s\,ds$ around this path should equal 4π times the current traversing any area of which this path is a boundary. If we consider the plane area which is shaded, we find simply $-4\pi i$ for this m.m.f. If, on the other hand, we construct a hemispherical surface, as shown, with the same perimeter, we find the magnetomotive force equal to *zero*, since no current crosses this surface. Hence Ampère's circuital law cannot be correct for the nonsteady state, and we are confronted with the problem of generalizing this law so that it will always be valid.

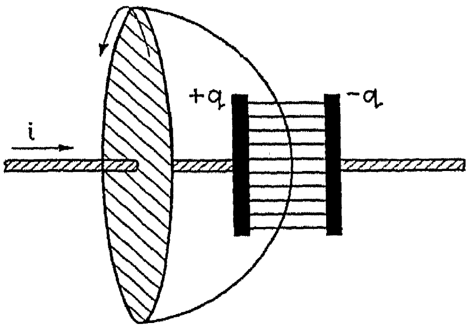

Fig. 98.

A satisfactory solution of this problem was first given by Maxwell, and it consists of generalizing the concept of electric current to include not only currents due to moving charges but also a new kind of current, called *displacement current*, which occurs whenever an electric field (or more precisely, the electric displacement vector D) varies with time. We can obtain the correct form for this current by considering the circuit of Fig. 98. If we set up an expression for a displacement current between the condenser plates just equal to the current in the wire, the Ampère circuital law can then be retained, and it gives a unique answer for the magnetomotive force around a closed path. In other words, we must postulate a displacement current such that the net current (both convection and displacement) leaving any *closed* surface is zero. This will be so if the displacement current emerging from the curved surface in Fig. 98 just equals the current entering across the plane shaded area. The latter has a magnitude dq/dt according to the equation of continuity, and this can be written as

$$\frac{dq}{dt} = \quad . \, d\sigma$$

where A is the area of the plate and σ the surface density of

charge. Since for the parallel-plate condenser we have $D = 4\pi\sigma$, we have for the displacement current between the condenser plates

$$i_d = \frac{1}{4\pi} A \frac{\partial D}{\partial t} \tag{4}$$

and a corresponding current density

$$j_d = \frac{1}{4\pi} \frac{\partial D}{\partial t} \tag{5}$$

Equation (5) is now postulated to be true in general; consequently, whenever an electric field changes with time, a current flows as given by this equation and a magnetic field is produced just as if a current due to moving charges existed. This remarkable discovery of Maxwell gives a symmetrical form to the fundamental laws of electromagnetism. The Faraday induction law implies that changing magnetic fields produce electric fields, and the Maxwell assumption implies that changing electric fields produce magnetic fields.

We can now reformulate the Ampère circuital law so that it holds in all cases, for both open and closed circuits. The total current, convection plus displacement, crossing any closed surface is zero; hence we may say that all distinction between open and closed circuits is lost. Writing the equation of continuity Eq. (2) in the form

$$\int_{\substack{\text{closed}\\ \text{surface}}} j_n \, dS + \frac{\partial q}{\partial t} = 0 \tag{6}$$

we can use Gauss's theorem for the flux of the electric displacement vector across this closed surface and have

$$\frac{1}{4\pi} \int_{\substack{\text{closed}\\ \text{surface}}} D_n \, dS = q$$

where q is the charge inside the surface. Substituting this value of q in Eq. (6), there follows

$$\int_{\substack{\text{closed}\\ \text{surface}}} \left(j_n + \frac{1}{4\pi} \frac{\partial D_n}{\partial t} \right) dS = 0 \tag{7}$$

or

$$\int_{\substack{\text{closed}\\ \text{surface}}} (j_n + j_{d_n})\, dS = 0 \tag{8}$$

and the form of the Ampère circuital law remains unchanged if we replace the ordinary current i by the sum of this current and the displacement current. Thus we find

$$\oint H_s\, ds = 4\pi i + \int \frac{\partial D_n}{\partial t}\, dS \tag{9}$$

This is one of the fundamental equations of electromagnetic theory.

Equation (9) can be used as it stands in any system of units in which all quantities are expressed in that system, as, for example, m.k.s. units. If we wish to use the Gaussian mixed system, however, expressing H in e.m.u., i and D in e.s.u., Eq. (9) takes the form

$$\oint H_s\, ds = \frac{4\pi}{c}\left(i + \frac{1}{4\pi}\int \frac{\partial D_n}{\partial t} dS\right) \tag{9a}$$

where

$$c = 3 \times 10^{10} \text{ cm./sec.}$$

Note that the displacement current flowing across any surface is just equal to $1/4\pi$ times the rate at which the flux of D across that surface changes with time. Thus, in Eq. (4), the displacement current flowing between the condenser plates is just $1/4\pi$ times the rate of change of the flux of $D(DA)$ leaving either plate. In ordinary conducting bodies the contribution of the displacement current to the total current flowing is completely negligible at low frequencies. To see this, let us imagine that we have an alternating-current density $j = J \sin 2\pi nt$ in a medium of conductivity σ. By Ohm's law we have $j = \sigma\mathcal{E}$ or $\mathcal{E} = j/\sigma$. Hence there will be a displacement current density

$$j_d = \frac{1}{4\pi}\frac{\partial D}{\partial t} = \frac{\epsilon}{4\pi}\frac{\partial \mathcal{E}}{\partial t} = \frac{\epsilon}{4\pi\sigma}\frac{\partial j}{\partial t}$$

or

$$j_d = \frac{\epsilon n}{2\sigma} J \cos 2\pi nt$$

and the ratio of the r.m.s. values of displacement to conduction current is $\epsilon n/2\sigma$.

We have no information concerning the value of ϵ for metals, so let us assume that it is equal to ϵ_0 for the moment. Using e.s.u., the ratio of displacement to ordinary current becomes of the order of magnitude n/σ. Now, since σ is of the order of 10^{17} per second for good conductors, we see that, even if ϵ were much larger than unity, the displacement current can be neglected in conductors for frequencies up to optical frequencies. Even at ultra high radio frequencies there is no error introduced by taking the conduction current to be the total current. In non-conducting bodies, however, this is not the case, and in empty space the displacement current is the total current. This discussion justifies the procedure employed in Chap. VII.

43. Plane Electromagnetic Waves in Vacuum.—Now let us examine some of the consequences of the concept of displacement current. We treat the case of empty space for which no convection currents or charges exist. The fundamental electromagnetic laws can be written for this case in the form:

$$\oint H_s\,ds = \int \frac{\partial D_n}{\partial t} dS \tag{10}$$

$$\oint \mathcal{E}_s\,ds = -\int \frac{\partial B_n}{\partial t} dS \tag{11}$$

$$\int\limits_{\substack{\text{closed}\\\text{surface}}} D_n\,dS = 0 \tag{12}$$

$$\int\limits_{\substack{\text{closed}\\\text{surface}}} B_n\,dS = 0 \tag{13}$$

If Gaussian units are employed, Eqs. (12) and (13) remain as they stand, but Eqs. (10) and (11) take the form

$$\oint H_s\,ds = \frac{1}{c}\int \frac{\partial D_n}{\partial t} dS \tag{10a}$$

$$\oint \mathcal{E}_s\,ds = -\frac{1}{c}\int \frac{\partial B_n}{\partial t} dS \tag{11a}$$

A simple and important solution of these equations may be obtained for which the electric and magnetic field vectors depend

on only one coordinate, let us say x, and on the time. For this case Eqs. (12) and (13) require that both the electric and magnetic vectors be normal to the x-axis, *i.e.*, that the x-components of all the vectors be zero. To see this, let us apply Eq. (12) at a given instant of time to the surface of an infinitesimal cube of sides dx, dy, and dz (Fig. 99). Equation (12) states that the total flux of D across the faces of this cube is zero. Since the components of D do not vary with y or z, the flux entering the volume across one of the faces $dx\,dy$, or $dx\,dz$, is just equal to the flux leaving across the corresponding opposite face. For the faces $dy\,dz$ the flux leaving through the right-hand face is

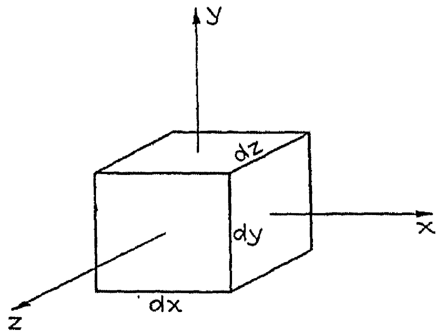

FIG. 99.

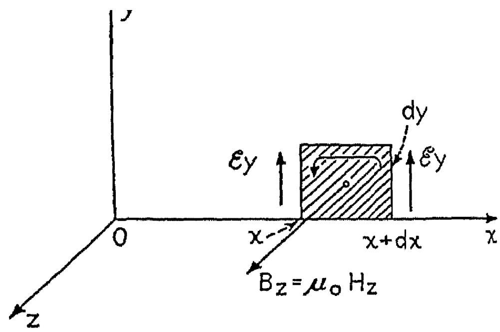

FIG. 100.

$D_x\,dy\,dz$, where D_x is the value of this component at the point where this surface is located. Similarly the flux leaving across the left-hand face is $-D_x\,dy\,dz$, where now D_x is taken at the point where this face is located. (The outward normal at this face is $-x$.) Since the sum of these terms must vanish in accordance with Eq. (12), it follows that D_x must have the same value at both faces and hence cannot vary with x. A similar proof holds for the magnetic induction B.

Let us start with the simplest assumption, namely, that $\mathcal{E}$ has only one component, $\mathcal{E}_y$, which may depend on x and on t. We now apply Eq. (11) to the elementary circuit shown in Fig. 100, proceeding in a counterclockwise direction as shown. The horizontal portions yield no contribution to the e.m.f. since $\mathcal{E}_s$ is zero for these sides. For the right-hand vertical side we have

$$\int \mathcal{E}_s\,ds = (\mathcal{E}_y)_{x+dx}\,dy$$

and for the left-hand side the corresponding term is

$$\int \mathcal{E}_s \, ds = -(\mathcal{E}_y)_x \, dy$$

Thus we obtain for Eq. (11)

$$\oint \mathcal{E}_s \, ds = [(\mathcal{E}_y)_{x+dx} - (\mathcal{E}_y)_x] \, dy = -\frac{\partial B_z}{\partial t} dx \, dy \tag{14}$$

Now $(\mathcal{E}_y)_{x+dx} - (\mathcal{E}_y)_x$ is just the change in $\mathcal{E}_y$ between the points x and $x + dx$ and can be written as $(\partial \mathcal{E}_y/\partial x) \, dx$. Substituting this value in Eq. (14), there follows

$$\frac{\partial \mathcal{E}_y}{\partial x} = -\frac{\partial B_z}{\partial t} = -\mu_0 \frac{\partial H_z}{\partial t} \tag{15}$$

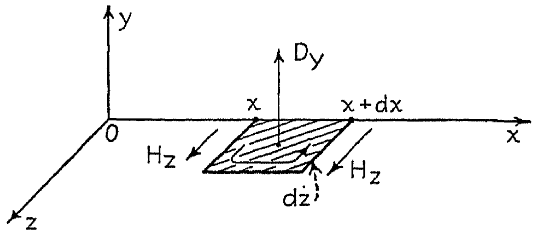

FIG. 101.

Next we calculate the magnetomotive force according to Eq. (10) around the elementary circuit shown in Fig. 101, proceeding in the direction indicated. We find exactly as in the preceding calculation

$$\oint H_s \, ds = [(H_z)_x - (H_z)_{x+dx}] \, dz = -\frac{\partial H_z}{\partial x} dx \, dz$$

and from Eq. (10) this must equal

$$\int \frac{\partial D_n}{\partial t} dS = \frac{\partial D_y}{\partial t} dx \, dz$$

There then follows

$$\frac{\partial H_z}{\partial x} = -\frac{\partial D_y}{\partial t} = -\epsilon_0 \frac{\partial \mathcal{E}_y}{\partial t} \tag{16}$$

The simultaneous solution of Eqs. (15) and (16) will then yield the electric and magnetic fields. We already see that, in order to have fields of the type we are investigating, the electric vector is accompanied by a magnetic vector perpendicular to it (in our case, B_z). If we differentiate Eq. (15) with respect to x, we obtain

$$\frac{\partial^2 \mathcal{E}_y}{\partial x^2} = -\mu_0 \frac{\partial^2 H_z}{\partial t \, \partial x}$$

and differentiating Eq. (16) with respect to t,

$$\frac{\partial^2 H_z}{\partial x \, \partial t} = -\epsilon_0 \frac{\partial^2 \mathcal{E}_y}{\partial t^2}$$

from which we obtain immediately

$$\frac{\partial^2 \mathcal{E}_y}{\partial x^2} = \epsilon_0 \mu_0 \frac{\partial^2 \mathcal{E}_y}{\partial t^2} \tag{17}$$

One can easily show that the other magnetic and electric vectors, H_z, B_z, and D_y satisfy exactly the same equation. Had we carried through our derivation using Eqs. (10a) and (11a) so that Gaussian units might be used, we would have in place of Eq. (17)

$$\frac{\partial^2 \mathcal{E}_y}{\partial x^2} = \frac{1}{c^2} \frac{\partial^2 \mathcal{E}_y}{\partial t^2} \tag{17a}$$

Equation (17) or (17a) is the equation for wave motion, the so-called wave equation, which we have encountered in our study of elastic vibrations, and the solutions represent waves traveling along the x-axis with a velocity $c = 1/\sqrt{\epsilon_0 \mu_0}$. Here we encounter a direct consequence of the assumption of a displacement current, the prediction of the existence of *electromagnetic waves* traveling with a velocity c which can be predicted from purely electrical measurements. The fact that the conversion factor c had a value nearly 3×10^{10} cm./sec., which is the velocity of light, led Maxwell to propose an electromagnetic theory of light. Nowadays the existence and observation of electromagnetic waves have become commonplace, but in Maxwell's time this prediction provided a critical test for his theory.

We see from our derivation that only transverse waves are predicted, both the electric and magnetic vectors being perpendicular to the direction of propagation. Let us consider a traveling sinusoidal wave which satisfies Eq. (17a). We can write for $\mathcal{E}_y$

$$\mathcal{E}_y = \mathcal{E}_0 \sin \omega \left(t - \frac{x}{c} \right) \tag{18}$$

representing a wave of amplitude $\mathcal{E}_0$ traveling in the positive x-direction with frequency $\nu = \omega/2\pi$ and velocity c. The wave length of the wave is given in the usual manner by

$$\lambda = \frac{c}{\nu} \tag{19}$$

The surfaces of constant phase are planes normal to the x-axis, given at any instant of time by the equation $x =$ constant; these

surfaces travel in the positive x-direction with a velocity c. The waves described by Eq. (18) are called *plane waves*, and the velocity c is often referred to as the *phase velocity* of these waves. At any fixed point of space the electric vector performs simple harmonic motion along a fixed direction, the y-axis, and the magnetic vector a similar vibration in a direction normal to this, the z-axis. Such a wave is called a *linearly polarized* wave, since the electric vector at any point has a fixed direction at all instants of time. More generally, we would have both y- and z-components of $\mathcal{E}$ of arbitrary amplitudes and phases but of the same frequency. Thus the resultant vector $\mathcal{E}$ at a fixed point of space would in general vary in such a manner that it would sweep out an ellipse, the superposition of two mutually orthogonal simple harmonic motions of arbitrary amplitudes and phase difference. This is shown diagrammatically in Fig. 102. Such a wave is called elliptically polarized, and the circularly polarized wave is the special case of the ellipse with equal major and minor axes.

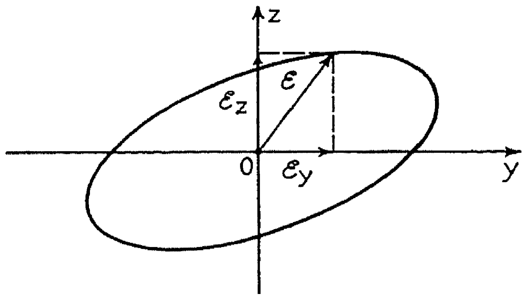

FIG. 102.

We must now investigate the relations between the amplitudes and phases of the electric and magnetic vectors in a plane wave. Suppose we start with Eq. (18) for the electric vector of a linearly polarized wave,

$$\mathcal{E}_y = \mathcal{E}_0 \sin \omega\left(t - \frac{x}{c}\right)$$

and differentiate this with respect to the time t. There follows

$$\frac{\partial \mathcal{E}_y}{\partial t} = \omega \mathcal{E}_0 \cos \omega\left(t - \frac{x}{c}\right)$$

Substituting this in Eq. (16), we find

$$\frac{\partial H_z}{\partial x} = -\omega\epsilon_0\mathcal{E}_0 \cos \omega\left(t - \frac{x}{c}\right)$$

and integrating

$$H_z = c\epsilon_0\mathcal{E}_0 \sin \omega\left(t - \frac{x}{c}\right)$$

Since $c = 1/\sqrt{\epsilon_0\mu_0}$ we can write this in the form

$$H_z = \sqrt{\frac{\epsilon_0}{\mu_0}}\mathcal{E}_0 \sin \omega\left(t - \frac{x}{c}\right)$$

showing that the magnetic vector oscillates in phase with the electric vector and the amplitudes of the electric and magnetic intensities are related by

$$\sqrt{\mu_0}H_0 = \sqrt{\epsilon_0}\mathcal{E}_0 \tag{20}$$

This relation is valid even in the Gaussian mixed system of units, as can be verified by using Eq. (16) in the form

$$\frac{\partial H_z}{\partial x} = -\frac{1}{c}\frac{\partial \mathcal{E}_y}{\partial t}$$

In a plane electromagnetic wave in empty space the electric and magnetic intensities are equal if the former is expressed in e.s.u.

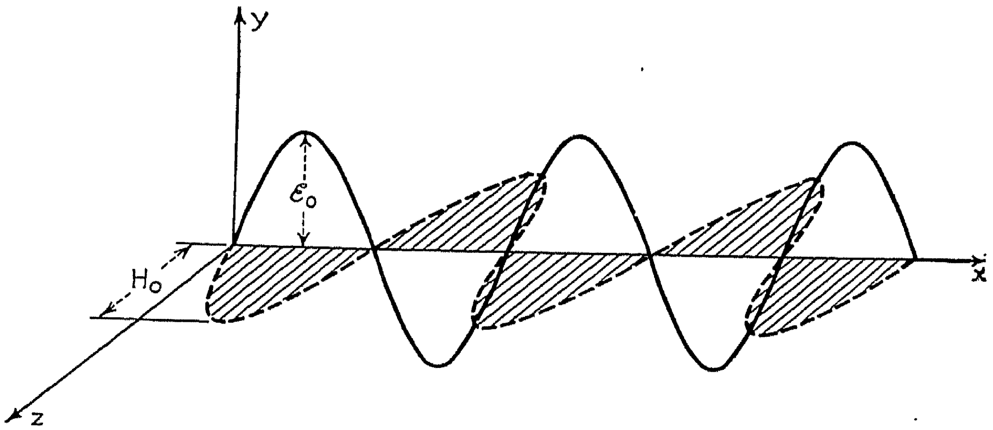

FIG. 103.

and the latter in e.m.u. In Fig. 103 are shown the space variations of $\mathcal{E}$ and H in a traveling linearly polarized sinusoidal wave.

44. Intensity and the Poynting Vector.—Electromagnetic waves such as we have just encountered carry energy, and this propagated energy is usually observed by absorption, *i.e.*, allowing the wave to impinge on some material surface and observing the changes which occur in the absorbing material. If we consider electromagnetic waves in empty space, the law of conservation of energy requires that the total energy flowing out of a fixed volume per unit time must just equal the rate of decrease of electromagnetic energy inside the volume. To formulate this

law quantitatively, we must introduce the concept of the *intensity* of electromagnetic radiation. We define the intensity S as the energy per unit area traversing an elementary surface per unit time when the elementary area is chosen normal to the direction of energy flow. From this definition we see that the intensity S is a vector, the magnitude of which is measured in ergs per square centimeter per second when e.s.u. or e.m.u. are employed, or in watts per square meter when m.k.s. units are utilized.

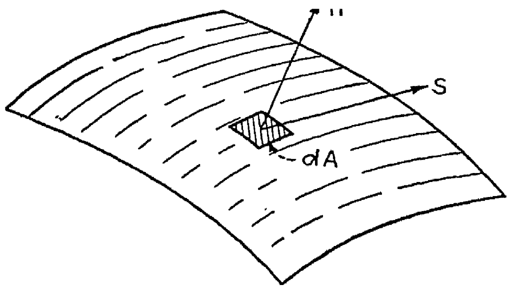

FIG. 104.

The rate at which energy flows across an arbitrary surface, open or closed, may be obtained as follows: Consider an element of area dA as shown in Fig. 104. The energy crossing this area per unit time is $S_n\, dA$, where S_n is the component of S along the normal n to the surface (Fig. 104). The total energy crossing the whole area per unit time is accordingly

$$\frac{dU}{dt} = \int S_n\, dA \tag{21}$$

where the integration extends over the whole surface. If the surface is closed, we can write for the electromagnetic energy inside the volume of space enclosed by this surface

$$U = \frac{1}{8\pi}\int (\epsilon_0 \mathcal{E}^2 + \mu_0 H^2)\, dv \tag{22}$$

The terms inside the parentheses are essentially the density of electric and magnetic energies at the point where the volume element dv is located. The rate of decrease of this expression with time must just equal the expression given in Eq. (21). Remembering that the surface and enclosed volume are fixed and do not vary with t, we have

$$\int_{\substack{\text{closed}\\ \text{surface}}} S_n\, dA = -\frac{1}{4\pi}\int \left(\epsilon_0 \mathcal{E}\frac{\partial \mathcal{E}}{\partial t} + \mu_0 H \frac{\partial H}{\partial t}\right) dv \tag{23}$$

As already pointed out, this equation is merely a statement of the conservation of energy as applied to electromagnetic waves in

empty space. From it we can derive the appropriate form for the intensity vector S in terms of the electric and magnetic field vectors. This intensity vector S is known as the *Poynting vector* after its discoverer. Let us apply Eq. (23) to the case of the linearly polarized plane waves which we discussed in the last section. Since the equality expressed by this equation must hold for an arbitrary volume, we shall choose an infinitesimal cube of sides dx, dy, and dz, as shown in Fig. 105.

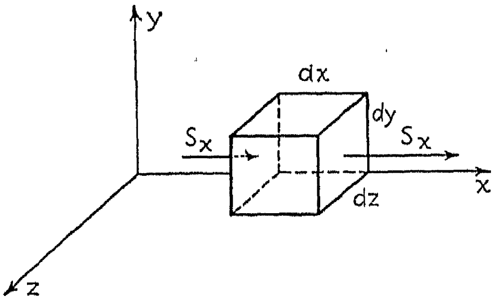

FIG. 105.

The right-hand side of Eq. (23) becomes, with the help of Eqs. (15) and (16),

$$-\frac{1}{4\pi}\left[\epsilon_0\mathcal{E}_y\left(-\frac{1}{\epsilon_0}\frac{\partial H_z}{\partial x}\right) + \mu_0 H_z\left(-\frac{1}{\mu_0}\frac{\partial \mathcal{E}_y}{\partial x}\right)\right] dx\,dy\,dz = \frac{1}{4\pi}\left(\mathcal{E}_y\frac{\partial H_z}{\partial x} + H_z\frac{\partial \mathcal{E}_y}{\partial x}\right) dx\,dy\,dz$$

or

$$\frac{1}{4\pi}\frac{\partial}{\partial x}(\mathcal{E}_y H_z)\,dx\,dy\,dz \tag{24}$$

Now consider the left-hand side of Eq. (23). This becomes $(S_x\,dy\,dz)_{x+dx} - (S_x\,dy\,dz)_x +$ similar terms in S_y and S_z. The first two terms represent the excess energy per unit time leaving the volume across the right-hand face $dy\,dz$ over that entering across the left-hand face $dy\,dz$. These two terms can be written in the form

$$\frac{\partial S_x}{\partial x}dx\,dy\,dz \tag{25}$$

This is just the same sort of expression as we have obtained in (24) and shows that in this case the Poynting vector is directed along the x-axis (the direction of propagation) and is given by

$$S_x = \frac{1}{4\pi}\mathcal{E}_y H_z \tag{26}$$

This is just the vector product of $\overrightarrow{\mathcal{E}}$ and $\overrightarrow{H}$, so that we write, in general,

$$\overrightarrow{S} = \frac{1}{4\pi}(\overrightarrow{\mathcal{E}} \times \overrightarrow{H}) \tag{27}$$

This is the correct general form for the Poynting vector and shows that the energy flows in a direction normal to the plane of $\mathcal{E}$ and H. If Gaussian units are employed, the corresponding expression for the intensity vector is

$$\vec{S} = \frac{c}{4\pi}(\vec{\mathcal{E}} \times \vec{H}) \tag{27a}$$

In the special case of plane waves where Eq. (20) gives the relation between the magnitudes of $\mathcal{E}$ and H, we obtain for the magnitude of the vector S

$$S = \frac{1}{4\pi}\sqrt{\frac{\epsilon_0}{\mu_0}}\mathcal{E}_0^2 \sin^2 \omega\left(t - \frac{x}{c}\right) = \frac{c}{4\pi}\epsilon_0\mathcal{E}_0^2 \sin^2 \omega\left(t - \frac{x}{c}\right) \tag{28}$$

where we have set

$$c = \frac{1}{\sqrt{\epsilon_0\mu_0}}$$

This shows that the intensity varies with the time but that the energy flow is always in the same direction, along the positive x-axis. The average rate of energy flow per unit area is

$$\overline{S} = \frac{c}{8\pi}\epsilon_0\mathcal{E}_0^2 \tag{29}$$

since the average value of $\sin^2 \omega\left(t - \frac{x}{c}\right)$ is $\frac{1}{2}$.

Equation (28) has a very simple physical meaning. We write it in the form

$$S = \frac{c}{4\pi}\epsilon_0\mathcal{E}^2 = \frac{c}{8\pi}(\epsilon_0\mathcal{E}^2 + \mu_0 H^2) \tag{30}$$

since $\epsilon_0\mathcal{E}^2 = \mu_0 H^2$ by Eq. (20). This states that the intensity in the wave at any point is equal to the velocity of the wave, c, times the energy density $\left[\frac{1}{8\pi}(\epsilon_0\mathcal{E}^2 + \mu_0 H^2)\right]$ at that point. If we consider a unit area normal to the direction of energy flow, the energy crossing this in time dt would be $S\,dt$, and this would occupy a volume of base unity and altitude $c\,dt$. Thus we expect an energy density equal to $\frac{S\,dt}{c\,dt}$, or S/c as is given by Eq. (30).

For convenience in numerical calculations, Eq. (29), giving the average intensity in a plane wave, can be written as

$\overline{S}$ (in watts per square meter) $= 1.32 \times 10^{-3}\mathcal{E}_0^2 = 2.65 \times 10^{-3}\mathcal{E}_{\text{r.m.s.}}^2$

where $\mathcal{E}_0$ and $\mathcal{E}_{\text{r.m.s.}}$ are expressed in volts per meter. In the Gaussian system one obtains

$\overline{S}$ (in ergs per square centimeter per second) $= 1.19 \times 10^9\mathcal{E}_0^2 = 2.38 \times 10^9\mathcal{E}_{\text{r.m.s}}^2$

with $\mathcal{E}_0$ and $\mathcal{E}_{\text{r.m.s.}}$ expressed in statvolts per centimeter.

Problems

1. Starting from the equation of continuity in the form

$$\int j_n\, dS = -\int \frac{\partial \rho}{\partial t} dv$$

show for a conducting medium obeying Ohm's law ($j = \sigma\mathcal{E}$) that in any volume element of the medium the following relation is valid:

$$\left(\frac{\partial \rho}{\partial t} + \frac{4\pi\sigma}{\epsilon_0}\rho\right) dv = 0$$

Use Gauss's law for the flux of D emerging from this volume. From the above relation show that, if initially one has a charge density ρ_0 at any point inside the conductor, the charge density ρ at any later time is given by

$$\rho = \rho_0 e^{-\frac{4\pi\sigma}{\epsilon_0}t}$$

The time constant $\epsilon_0/4\pi\sigma$ is known as the relaxation time.

2. A parallel-plate condenser having circular plates, each of area A, is connected to an a.c. generator so that the charge on the plates is given by $q = q_0 \sin \omega t$. Neglecting end effects, the lines of H are circles whose centers lie on the axis of symmetry of the system. Prove that the magnetic intensity at any point between the plates is given by

$$H = \frac{2\pi r\omega}{A} q_0 \cos \omega t$$

where r is the distance from the axis to the field point.

3. For the arrangement of Prob. 2 derive an expression for the flux of B in the space between the condenser plates.

4. Compute the frequencies of electromagnetic waves of the following wave lengths: 10^7 cm. (audio frequencies); 10^4 cm. (radio frequencies); 10 microns (heat waves) [1 micron $= 10^{-3}$ mm.]; 5,000 angstrom units (optical waves) [1 angstrom $= 10^{-8}$ cm.]; 10^{-1} angstrom (X rays).

5. Following the method outlined in the text, show that, for a plane wave traveling in the x-direction, the electric and magnetic intensities satisfy the relations

$$\frac{\partial \mathcal{E}_y}{\partial x} = -\mu_0 \frac{\partial H_z}{\partial t}; \qquad \frac{\partial \mathcal{E}_z}{\partial x} = \mu_0 \frac{\partial H_y}{\partial t}$$

$$\frac{\partial H_y}{\partial x} = \epsilon_0 \frac{\partial \mathcal{E}_z}{\partial t}; \qquad \frac{\partial H_z}{\partial x} = -\epsilon_0 \frac{\partial \mathcal{E}_y}{\partial t}$$

From these equations show that $\mathcal{E}_y$ and $\mathcal{E}_z$ separately satisfy the wave equation (Eq. 17 of the text).

6. The average rate at which the earth receives radiant energy from the sun at noon is 2.2 cal./min. per cm.^{2} Compute the r.m.s. values of the electric intensity and of the magnetic induction in sunlight at the earth's surface.

7. A 100-watt lamp radiates all the energy supplied to it uniformly in all directions. Compute the r.m.s. values of the electric and magnetic vectors at a point 1 meter from the lamp. What is the energy density of electromagnetic radiation in ergs per cubic centimeter at this point?

8. Two linearly polarized plane waves traveling in the same direction (along the x-axis) have their electric vectors in the same direction (the y-axis) but have different amplitudes and frequencies. Prove that the average intensity of the resultant wave is equal to the sum of the average intensities of each wave. (Average over a large number of periods.) From this result show that the r.m.s. value of $\mathcal{E}$ for the resultant wave is equal to the square root of the sum of the squares of the r.m.s. values of $\mathcal{E}$ for the individual waves.

9. A plane electromagnetic wave with a maximum value of the electric intensity of 0.01 volt/cm. falls normally on a surface which is perfectly absorbing. The mass of the surface is 10^{-3} gram/cm.^{2}, and its specific heat is 0.2. What is the rate of increase of temperature of the surface?

CHAPTER IX

RADIATION OF ELECTROMAGNETIC WAVES

In the preceding chapter we have seen how the introduction of the concept of displacement current by Maxwell led to the prediction of the possibility of free electromagnetic waves in empty space. We now turn to the question of the sources of electromagnetic waves, *i.e.*, to the methods of producing waves of different wave lengths. The electromagnetic spectrum is most conveniently classified according to the sources utilized for the production of the radiation of different wave lengths and in a broad sense falls naturally into two large divisions: First—and this is the region which will occupy our attention in this chapter—we have the *long wave* region, comprising waves of wave length longer than about 1 mm. in empty space; and, secondly, we have the region from about 1-mm. wave length down to the shortest waves known, at present about 10^{-11}-cm. wave length. The fundamental difference between these two regions lies in the so-called *coherence* of the radiation. In the short wave-length region the electromagnetic radiation field consists of the superposed effects of a huge number of elementary wave trains, each of atomic origin; one cannot control the relative phases and the lengths of these elementary wave trains. Such radiation, for which the relative phases of the individual elementary waves bear no fixed relations to each other and hence for which there is a completely random distribution of these relative phases, is called *incoherent*. In the long wave-length region, on the other hand, the radiation can be produced by large-scale generators or oscillators and is *coherent;* amplitude, phase, and length of the wave trains can be controlled and maintained with fixed relative values. The difference between incoherent and coherent radiation is quite similar to the difference between the motion of a large number of molecules (in kinetic theory), in which the positions and velocities of the individual molecules are completely random and uncontrollable, and the motion of one or more large-scale particles, in which the individual motions may be controlled.

We defer the discussion of atomic radiation (heat, light, X rays, etc.) to later chapters and confine ourselves in this chapter to a treatment of the long wave-length region. Here we shall concern ourselves principally with plane electromagnetic waves of the sort studied in the preceding chapter and with *spherical* waves, waves for which the constant-phase surfaces are concentric spherical surfaces. It should be pointed out that the term spherical wave does not necessarily imply spherical symmetry in the distribution of amplitude of the wave over a surface of constant phase—the amplitude may vary in a complicated manner from point to point of such a spherical surface—but simply that these constant-phase surfaces are spherical.

45. Electromagnetic Waves on Wires.—In a traveling electromagnetic wave the electric and magnetic vectors vary with both position and time, the changing magnetic vector inducing an electric field and the changing electric vector (the displacement current) inducing a magnetic field. Since stationary charges produce electrostatic fields and since charges moving with constant velocity (steady currents) produce steady magnetic fields, it is evident that electromagnetic wave fields can be produced by accelerated charges or by currents varying with the time. The simplest sort of accelerated motion of charges or variation of current with time is that involving simple harmonic motion, and we have already studied this sort of oscillation in the case of a simple circuit consisting of an inductance and a capacitance in series. Neglecting the resistance in the circuit, the conduction current in such a circuit will vary sinusoidally (as will the charges on the condenser plates) with a frequency ν given by

$$= \frac{1}{2\pi}\sqrt{\frac{1}{LC}} \tag{1}$$

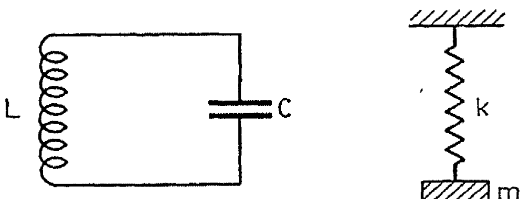

FIG. 106.

where L is the inductance and C the capacitance [Fig. 106 and compare Eq. (36), Chap. VII]. This circuit is the analogue of the mechanical system of a mass m moving under the action of a spring of stiffness coefficient k, the motion being simple harmonic with frequency

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}} \tag{1a}$$

In the mechanical motion there is a periodic interchange of the kinetic energy of the mass m (assuming it to be large enough compared to the mass of the spring so that we may neglect the latter) and the potential energy of the spring. We must stress the point that the mass m (the carrier of kinetic energy) and the spring (the carrier of potential energy) are completely separated. Similarly in the electric circuit, we have a periodic interchange of the electric energy stored in the condenser (the analogue of potential energy) and the magnetic energy (the analogue of kinetic energy) in the field of the inductance. Once again we have a complete separation of the electric and magnetic energies, neglecting the small magnetic field produced by the displacement current between the plates of the condenser and the small distributed capacity of the inductance and connecting wires. This separation of the "seats" of electric and magnetic energies becomes more and more complete, the larger the inductance L and the larger the condenser C. Since an electromagnetic wave involves the overlapping of the fields of $\mathcal{E}$ and H, the possibility of production of waves from the circuit of Fig. 106 is practically nil, just as one cannot obtain mechanical waves in the case of the spring and mass of the same figure.

In mechanics we have seen that wave motion could be set up in extended continuous bodies in which mass and restoring forces are continuously distributed throughout the media. Such is the case of a stretched string or of an air column in a tube. In these cases we may no longer think of the potential energy being restricted to one region of space and kinetic energy to another, but there is a periodic variation of kinetic and potential energy at every point of the medium. Now exactly the same requirements must be met in the electrical problem if one is to set up wave motion. Our circuit must not have concentrated capacitance and inductance, but the configuration must be such that both the electric and magnetic fields overlap and exist in the same region of space. Perhaps the simplest circuit which fulfills this requirement is a transmission line, consisting of a pair of parallel wires, or the more symmetrical coaxial cable. Here we may consider the two conductors as forming the plates of an extended condenser, the electric field being largely in the region of space between them, and the currents flowing along these conductors as giving rise to a magnetic field in the same

region of space. We should expect that electromagnetic waves could be set up in this system, and we shall now show that this is true.

Consider a very long parallel pair of wires connected to an a.c. generator as shown in Fig. 107. We shall consider the case for which the resistance of the conductors is negligible, so that the charges on the wires flow along their surfaces. At a given

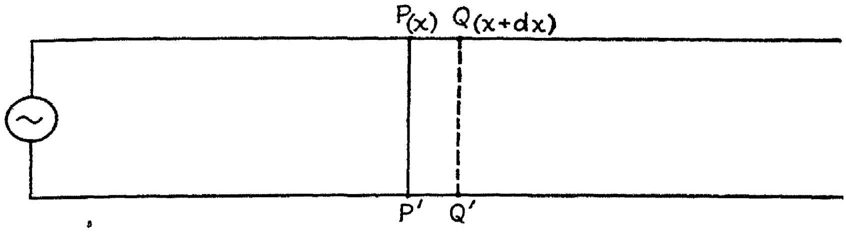

FIG. 107.

instant of time there will be a definite distribution of charge on the conductors, and this distribution will change with time, giving rise to a current distribution. If τ is the charge per unit length (linear charge density)—this varies from point to point and with the time—then the charge residing on the element of length $dx = PQ$ (Fig. 107) is $\tau\, dx$ at a definite instant of time with an equal and opposite charge on the element $P'Q'$. If C' is the capacity per unit length of the system, then the voltage E between P and P' (at the point x) is related to the charge by

$$\tau = C'E \tag{2}$$

Furthermore, the voltage drop $-dE = -\frac{\partial E}{\partial x}dx$ is given by

$$-\frac{\partial E}{\partial x}\, dx = L'\, dx\, \frac{\partial i}{\partial t} \tag{3}$$

where L' is the inductance per unit length. The charge density and current are related by the equation of continuity. For the case at hand we have for the decrease of charge on the element dx (PQ) in time dt

$$-dq = -\frac{\partial \tau}{\partial t}\, dx\, dt$$

and this must equal the difference of the currents at Q and P $\left(di = \frac{\partial i}{\partial x}dx\right)$ times the time dt. Thus we have

$$-\frac{\partial \tau}{\partial t}\, dt\, dx = \frac{\partial i}{\partial x}\, dx\, dt \tag{4}$$

If we differentiate Eq. (2) with respect to the time and use Eq. (4), we obtain

$$-\frac{\partial i}{\partial x} = C'\frac{\partial E}{\partial t} \tag{5}$$

and Eq. (3) may be written as

$$\frac{\partial E}{\partial x} = -L'\frac{\partial i}{\partial t} \tag{6}$$

Equations (5) and (6) determine the current and voltage distribution along the conductors. Differentiating Eq. (5) with respect to x and Eq. (6) with respect to t, we can eliminate E from these equations and find

$$\frac{\partial^2 i}{\partial x^2} = L'C'\frac{\partial^2 i}{\partial t^2} \tag{7}$$

which is the wave equation, showing that traveling waves of current (and charge) may exist in this line. Similarly from Eqs. (5) and (6) one obtains for E

$$\frac{\partial^2 E}{\partial x^2} = L'C'\frac{\partial^2 E}{\partial t^2} \tag{8}$$

showing that the voltage between conductors also obeys the wave equation. The velocity of these electromagnetic waves is given by

$$v = \frac{1}{\sqrt{L'C'}} \tag{9}$$

Now it is a remarkable fact that, for any pair of *parallel* straight conductors of uniform cross section, the product of the inductance per unit length and the capacitance per unit length is just $\epsilon_0\mu_0 = 1/c^2$; hence the velocity given by Eq. (9) is just the velocity of free electromagnetic waves in empty space. We can check this for the case of a coaxial cable. For this case Eq. (13) of Chap. III gives

$$C' = \frac{\epsilon_0}{2 \ln (b/a)}$$

whereas Eq. (16) of Chap. VI yields

$$L' = 2\,\mu_0 \ln (b/a)$$

If, for example, the generator at the end of the line of Fig. 107 generates a voltage $E_0 \sin 2\pi\nu t$, then in the steady state one obtains a traveling sinusoidal voltage wave along the line given by

$$E = E_0 \sin 2\pi\nu\left(t - \frac{x}{c}\right) \tag{10}$$

One has similar expressions for the traveling waves of current and charge which may be immediately obtained from Eqs. (5) and (6) when one utilizes Eq. (10) for E.

Thus far in our discussion of electromagnetic waves on a perfectly conducting transmission line we have employed the language of circuit theory, using the notions of capacity and inductance. Now let us examine the question in terms of the field theory, *i.e.*, in terms of the electric and magnetic fields in the space around the conductors, so that we may clearly see the connection between the circuit ideas and the plane electromagnetic waves in free space which were introduced in Chap. VIII. For this purpose it will be easier to visualize the state of affairs by imagining the wires replaced by flat strips without changing the essentials of the problem. For this case the lines of $\mathcal{E}$ will be parallel to each other, and, at a given instant of time, the magnitude of $\mathcal{E}$ will be constant at any fixed point along the line and equal to E/d, d being the separation of the conductors. This neglects fringing, and we must confine ourselves to points well within the boundaries. Similarly, near the central portion, the lines of B (or H) will be parallel to each other at right angles to $\mathcal{E}$ and to x and the magnitude of either will be essentially constant in this region. Thus we have the identical picture for the electromagnetic field between the conductors as we had for the case of a linearly polarized plane wave in free space, $\mathcal{E}$ having only a y-component and H a z-component. It now becomes clear that the wave equation for the voltage (Eq. 10) is identical with that for $\mathcal{E}_y = E/d$, which one derives directly from the fundamental electrodynamic laws for plane waves [compare Eq. (18) of Chap. VIII]. Thus we have traveling waves of $\mathcal{E}$ and H with the Poynting vector S directed along the x-axis (along the line) so that the energy flow is always in this direction. The state of affairs is illustrated schematically in Fig. 108 for a traveling sinusoidal wave at a given instant of time.

The vertical lines terminating on the conductors represent the lines of $\mathcal{E}$, the magnitude of $\mathcal{E}$ indicated by the separation of the lines. The dots indicate the distribution of the magnetic field vector H, the density of dots indicating roughly the manner in which the magnitude of this vector varies with position along the line. The arrows just above and below the $+$ and $-$ charges indicate the distribution of current in the conductors. To obtain the picture of the traveling sine wave, we must imagine the whole diagram moving uniformly to the right. The sine

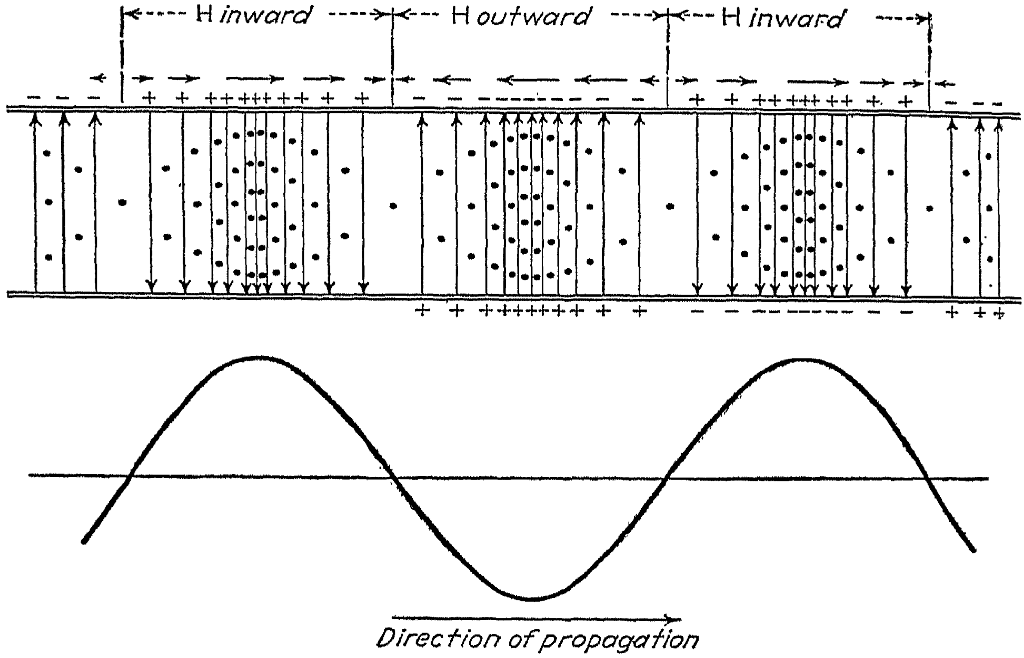

FIG. 108.

curve sketched below the figure represents the variation of either charge, electric intensity or magnetic field, current or voltage between wires, with distance along the line. Note that they are all in phase with one another. It must be emphasized that these results hold strictly only for infinitely good conductors for which the charges reside on the surface. When resistance is taken into account, the velocity of propagation is less than that of electromagnetic waves in empty space, and the fields are distorted, there being a component of the Poynting vector directed perpendicular to and into the wires. However, at very high frequencies due to skin effect, the current is concentrated very near the conductor surfaces so that the velocity of the waves approaches c and is independent of the conductor material.

Thus far we have been discussing waves traveling on an infinitely long transmission line. Just as in the case of mechanical waves, one obtains reflections of these waves at the end of the line when the latter is of finite length, and in some cases one may obtain *standing waves.* Thus, for example, suppose we imagine the line of Fig. 107 of length $l/2$, open at the far end as shown in Fig. 109. In this case we obtain very nearly complete reflection at the open end, and the superposition of the incident and reflected waves gives rise to standing waves just as in the case of mechanical waves. In Fig. 109 are sketched the first few modes of

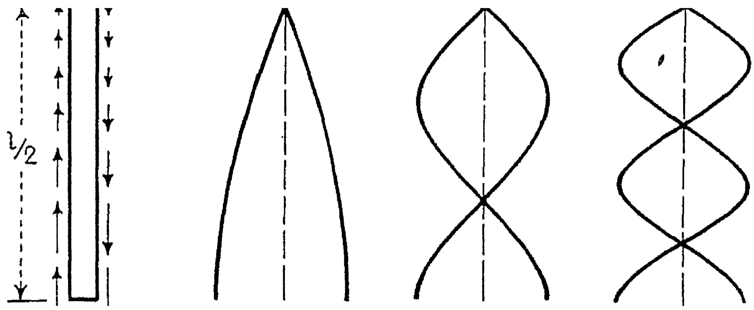

FIG. 109.

vibration of an open line, representing the distribution of current along the wires. The arrows schematically indicate the variation of current in both wires with position at a given instant of time for the fundamental mode of oscillation. The analogies with the case of vibrating mechanical systems are evident. In contrast to the case of traveling waves, the voltage and current are 90° out of phase at every point; consequently there is no propagation of energy on the average, merely an oscillation back and forth along the line. Thus there is a current node at the open end and a voltage maximum at this point, whereas at the short-circuited end there is a voltage node and a current maximum. One can readily show that the standing-wave pattern of the electric and magnetic fields is similar to that of voltage and current, the electric field having a node at the short-circuited end and an antinode at the open end, with just the converse behavior of the magnetic field. In the fundamental mode of vibration, the voltage between the wires varies according to the equation

$$E = E_m \sin 2\pi\nu t \sin \frac{\pi x}{l}$$

and from Eq. (5) we have

$$\frac{\partial i}{\partial x} = -E_m C' 2\pi\nu \cdot \cos 2\pi\nu t \cdot \sin \frac{\pi x}{l}$$

Integrating this with respect to x, the current distribution becomes

$$i = 2E_m C' l\nu \cdot \cos 2\pi\nu t \cdot \cos \frac{\pi x}{l} \qquad (12)$$

which brings out the fact that the 90° phase difference between voltage and current occurs both for the space distribution at a given instant of time and to the sinusoidal time variations at a given point on the line. Further details are left to the problems.

46. The Radiation Field of an Oscillating Dipole.—We have just seen how both traveling and standing electromagnetic waves may be produced by currents and charges on a pair of parallel conductors and must now inquire into the question as to how such waves may be disengaged from the conductors and become free electromagnetic waves. This is the problem of the radiation of waves from accelerated charges and is sufficiently complicated that we cannot undertake a quantitative analysis. We shall therefore arrive at the radiation formulas with the help of qualitative and semiquantitative reasoning.

Since the Maxwell displacement current is primarily responsible for the possibility of existence of free electromagnetic waves, let us consider more carefully the distribution of displacement current in the case of waves on a transmission line. In any case the lines of current flow (conduction and displacement current) must form closed curves. For the transmission line, the lines of displacement current terminate on the conductors, since they are coincident with the lines of $\mathcal{E}$, and there they join continuously with the lines of conduction current on the surfaces of the conductors. This is in sharp contrast to the situation present in the case of free electromagnetic waves in empty space where there is no conduction current and hence the lines of displacement current (and of $\mathcal{E}$) form closed curves. To understand how it is possible to produce free electromagnetic waves from currents or waves on conductors, we must see why the lines of $\mathcal{E}$ should break away from the conductors and form closed curves. The clue to the solution of this problem lies in the fact that electromagnetic disturbances are propagated with a finite velocity. If

we consider the electric and magnetic fields at a distance r from an oscillating charge distribution on an element of length of a conductor, we readily see that the fields oscillate essentially in phase with the charge and current variations, provided that the distance r is very much shorter than a wave length of waves of the same frequency. In this case the time required for a change in the charge and current distribution to produce a corresponding change in the fields is very short compared to the period of oscillation; consequently at any instant of time the fields are essentially the same as would be produced by a steady-state distribution of charge and current. On the other hand, if we consider the state of affairs at a distance large compared to a wave length, *i.e.*, such that the time of propagation of the fields from the source to the field point is much longer than one period of oscillation, then *retardation effects*, as they are termed, become important, the fields are distorted from the patterns obtained from fixed-charge distributions and steady currents, and this field distribution corresponds to that of traveling waves. Suppose we fix our attention on two field points, both lying on a straight line drawn through the source and both at distances from the source which are much greater than a wave length. Furthermore, let the separation of these points be just half a wave length. At a given instant of time there will be a definite direction of $\mathcal{E}$ and of the displacement current at the outer point, and evidently at the other field point the direction of these vectors must be just reversed, since the time required for the fields to propagate from one field point to the other is just half a period. In this time the charge and current distributions at the source have just reversed. Now the lines of $\mathcal{E}$ and $\partial\mathcal{E}/\partial t$ through these two points cannot conceivably both terminate on the conductor which is the source, and, since these lines cannot start or stop in empty space, they must close on each other, forming the closed lines of $\mathcal{E}$ necessary for free electromagnetic waves. In the region of space which is of the order of magnitude of a wave length distant from the source, the field patterns are exceedingly complex. This is the transition region between the two types of field just discussed.

When the electromagnetic waves travel along a pair of parallel wires, the fields are confined largely to the region of space between the conductors, making the geometrical arrangement unfavorable

for radiation. Even if the separation is made large compared to a wave length—and we have tacitly assumed in Sec. 45 that it was very small compared to a wave length—no appreciable advantage is gained, since lines of $\mathcal{E}$ starting at the surface of a conductor then terminate on other points of the same conductor rather than on the other wire and the wave is still guided, so to speak, by the wires. Let us, however, imagine that we deform the circuit shown in Fig. 109 by increasing the angle between the wires until they eventually form a straight conductor of length l (Fig. 110).

The electric field due to a fixed-charge distribution on this circuit evidently spreads out over a larger and larger region of

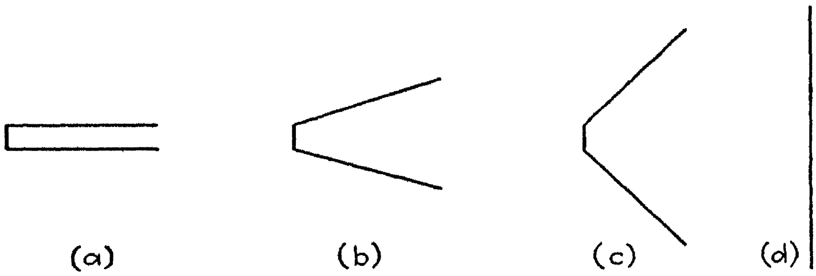

FIG. 110.

space as the angle between the original pair of parallel wires is increased. In Fig. 110*d* we have the straight wire (an antenna) with its field extending to large distances. This is the so-called *electric dipole* (more precisely, an extended dipole), and in it we are now to imagine, not a static distribution of charge, but a standing sinusoidal wave of charge and current, let us say in the fundamental mode of vibration corresponding to Fig. 109. At any point of this antenna, *i.e.*, in any length element dl, the current and charge will vary sinusoidally with the time, but there will be a progressive phase shift of these sinusoidal oscillations as we move the element dl along the antenna. Hence to calculate the radiation field of such a system, we must superpose the contributions from all the elementary lengths dl, taking into account the relative phases of these contributions. Thus the fundamental problem is that of studying the field of an element dl carrying an alternating current, and, as we shall show in a moment, this is equivalent to a tiny dipole oscillating in simple harmonic motion. An antenna of finite length evidently becomes equivalent to such a dipole if it is driven in *forced*

oscillation at frequencies much below its lowest natural frequency so that essentially no phase difference exists between the currents at different points of the conductor. In this case the length of the antenna is very small compared to a wave length and consequently extremely small compared to the distances from the antenna to the field points where the wave field exists.

We are now ready to discuss the nature of the radiation field of an oscillating dipole in more detail. The lines of B or H due to our dipole consist of circles with their centers on the axis of the dipole, just as in this case of a straight wire carrying current, and this is true at all distances from the dipole, although the variation of magnitude with distance is radically different at distant and near points. The distribution of the lines of electric intensity is very complicated, becoming relatively simple only at large distances. In this region the lines of $\mathcal{E}$ are perpendicular both to H and to the radius vector drawn from the dipole to the point P at which we are considering the field. This is illustrated in Fig. 111. H at the point P is directed outward from the plane of the paper. Furthermore, $\mathcal{E}$ and H are equal to each other (in Gaussian units) at the point P, just as in the case of plane waves. The Poynting vector S is directed outward along r so that we have energy traveling radially outward from the source at O and thus have spherical electromagnetic waves. Before we discuss the analytical expressions for $\mathcal{E}$ and H for these spherical waves, let us consider the physical nature of the field.

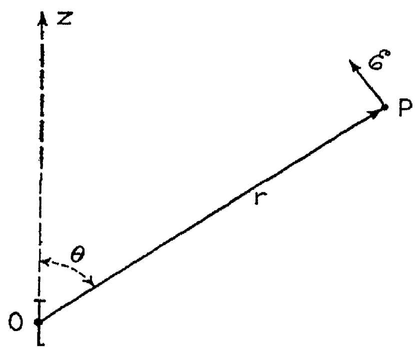

Fig. 111.

The oscillating charge and current in the antenna produce electric and magnetic fields which propagate outward from the source with the velocity of light. During a half cycle, fields build up in one direction, and then they reverse and decrease (build up in the opposite direction). These alternations of the field strengths lag behind the alternations of charge and current in the source, the lag increasing with increasing distance from the source because of the finite velocity of propagation of these fields. If there were no phase lag (infinite velocity of propagation), the average power flowing into the field would be zero

when averaged over a cycle, as much energy flowing out as returning during this time just as in the case of the field energy in a condenser at low frequencies. Because of the progressive phase lag with distance, however, there is not exact cancellation of outward and inward flow of power from the source, and there is a small amount radiated in the form of the spherical waves mentioned previously. It must not be thought that these waves exist only at large distances from the source. They are present in all space, even close to the source. The resultant fields, however, are to be thought of as the superposition of this wave field and other fields which do not represent traveling waves. The latter fields, the energy of which alternately leaves and returns to the source, are very large compared to the wave field near the antenna, but they decrease much more rapidly with distance from the source than the wave field, so that only at distances large compared to the wave length does the resultant field correspond to traveling spherical waves.

Now let us consider the variation of the magnitudes of $\mathcal{E}$ and H with position and time. Since $\mathcal{E}$ and H are equal in the wave field, it will be sufficient to consider one of them. First consider the variation with distance r from the source and with the time. We would expect that $\mathcal{E}$ would vary sinusoidally with the time with a frequency equal to that of the dipole but retarded in phase by an amount $2\pi r/\lambda = 2\pi\nu r/c$. Furthermore, the amplitude of this oscillating vector must be proportional to $1/r$, as we can immediately see from the conservation of energy: The total energy crossing a spherical surface of radius r per unit time must be independent of r, since there are no sources or sinks of energy in empty space. The Poynting vector, which is proportional to $\mathcal{E}^2$, must hence vary as $1/r^2$, so that, when it is integrated over the surface of the sphere (the area of which is $4\pi r^2$), the result will be independent of the position of the spherical surface. Thus we expect an expression for $\mathcal{E}$ (or H) of the form

$$\frac{\sin 2\pi\nu\left(t - \frac{r}{c}\right)}{} \tag{13}$$

We have been referring to the source either as an antenna or as a dipole. Let us examine the connection between the wire of length l of Fig. 110d with a standing wave of current and a

dipole. Let $p = qz$ be the dipole moment at any instant of time and $P = qz_{max}$ the maximum value thereof. Since for simple harmonic motion

$$z = z_{max} \cos 2\pi\nu t$$

we have

$$p = qz = P \cos 2\pi\nu t$$

Now a charge q moving with velocity v is equivalent to a current element $i \cdot dl$, so that we have

$$\frac{dp}{dt} = q\frac{dz}{dt} = - \qquad \sin$$

Thus an oscillating dipole is entirely equivalent to a sinusoidal current in a conductor element of length equal to the amplitude of motion of the dipole, and the amplitude of this current is equal to 2π times the product of the frequency and the charge of the dipole. Hence each element of length of the antenna (the extended dipole) is equivalent to an oscillating dipole.

p=qz

FIG. 112.

The magnitude of $\mathcal{E}$ depends on the angle θ (Fig. 111) between r and the direction of the dipole moment in just the same way as the field of a static dipole, being maximum for $\theta = \pi/2$, zero for $\theta = 0$, and proportional to $\sin \theta$ for intermediate values of θ. Thus the complete space dependence of $\mathcal{E}$ (and H) in the wave field is, using Eq. (13),

$$|\mathcal{E}| = |H| \qquad \sin 2\pi\nu\left(t - \right. \qquad \sin \theta \tag{14}$$

and there remains only the question of the proportionality constant. As we have already mentioned, electromagnetic waves are produced by accelerated charges (varying currents), since steady charges and currents give rise to static electric and magnetic fields, respectively. Thus we would expect the proportionality constant of Eq. (14) to be proportional to the acceleration of the oscillating dipole charge, *i.e.*, to the quantity $4\pi^2\nu^2P$, and this is exactly the result one obtains from the exact theory. We now write the complete expression for the magnitude of the electric or magnetic vector for the spherical wave emitted by an oscillating dipole. It is, in Gaussian units,

$$|\mathcal{E}| = |H| = \quad \sin\theta \cdot \sin \quad - \frac{r}{c}\Big) \tag{15}$$

and the directions of these vectors (at a given instant of time) are shown in Fig. 113. Since $\mathcal{E}$ and H both vary as $\sin\theta$, the intensity of the wave, S, is not spherically symmetrical but varies as $\sin^2\theta$, maximum radiation at right angles to the direction of the dipole moment and zero along the axis. The surfaces of constant phase are, however, concentric spherical surfaces.

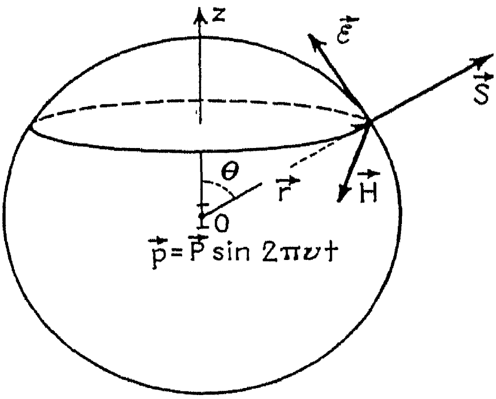

FIG. 113.

As a final step we shall compute the total rate of emission of electromagnetic energy from the dipole. The Poynting vector has the magnitude

$$|S| = \frac{c}{4\pi}\mathcal{E}^2 = \frac{4\pi^3\nu^4P^2}{c^3r^2}\sin^2\theta \cdot \sin^2 2\pi\nu\left(t - \frac{r}{c}\right) \tag{16}$$

and is normal to the constant-phase surfaces. Since the time average of $\sin^2 2\pi\nu\left(t - \frac{r}{c}\right)$ is $\frac{1}{2}$, we have for the average rate of energy flow per unit area across a sphere of radius r

$$\overline{S} = \frac{2\pi^3\nu^4P^2}{c^3r^2}\sin^2\theta \tag{17}$$

and to obtain the total rate of emission of energy (the radiated power), we must evaluate the integral $\int \overline{S} \cdot dA$ over the surface of a sphere of radius r. The appropriate element of area is a ring of width $r\,d\theta$ and circumference $2\pi r \sin\theta$ (Fig. 114), so that we have

$$\overline{\frac{dE}{dt}} = \int \overline{S}\,dA = \frac{4\pi^4\nu^4P^2}{c^3}\int_0^\pi \sin^3\theta\,d\theta \tag{18}$$

To evaluate the integral, we write

$$\sin^3\theta\,d\theta = (1 - \cos^2\theta)\sin\theta\,d\theta = -(1 - \cos^2\theta)d(\cos\theta)$$

so that

$$\int^{\pi} \sin^3 \theta \, d\theta = -\int_{\theta = 0}^{\theta = \pi} (1 - \cos^2 \theta) d(\cos \theta) = -\left[\cos \theta - \frac{\cos^3 \theta}{3}\right]_0^{\pi} = \frac{4}{3}$$

Substituting this value in Eq. (18), we obtain as the final result

$$\overline{\frac{dE}{dt}} = \frac{16\pi^4\nu^4}{3c^3} P^2 = \frac{\omega^4 P^2}{3c^3} \tag{19}$$

where $\omega = 2\pi\nu$.

Thus we see that the average rate of energy emission is proportional to the square of the amplitude of the dipole moment and to the fourth power of the frequency.

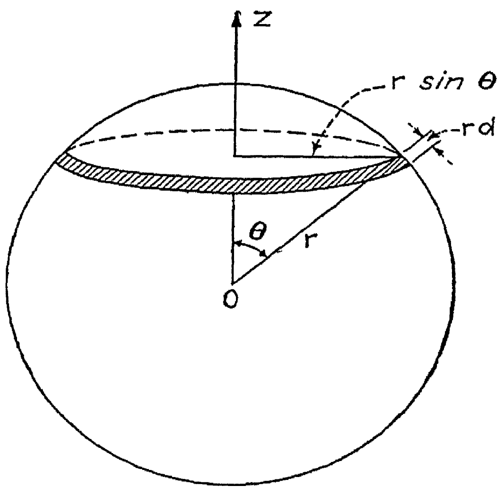

FIG. 114.

The extension of the foregoing results to the case of an antenna of finite length oscillating at one of its natural frequencies is a straightforward integration of the expressions for $\mathcal{E}$ and H over the elements of length of the antenna. We shall not carry it through, however, since we shall have occasion to solve a similar problem in optics later which will bring out all the essential points.

Problems

1. A voltage $E = E_0 \sin 2\pi\nu t$ is applied to one end of a very long transmission line which consists of a pair of parallel conductors. Neglecting resistance, show that the generator delivers a current i to the line which is related to the voltage by

$$E = \sqrt{\frac{L'}{C'}} \cdot i$$

where L' and C' are the inductance and capacitance per unit length of the line.

2. Suppose the transmission line of Prob. 1 is a coaxial cable, the central conductor having a radius a and the sheath an inner radius b. Assuming ideal conductors:

a. Compute expressions for $\mathcal{E}$ and H at any point between the conductors, and from these obtain the direction and magnitude of the Poynting vector.

b. Find the rate of energy flow perpendicular to a cross section of the cable by integrating the Poynting vector over such a cross section, and show that its average value equals the average power delivered by the generator.

3. Prove, for the coaxial cable of Prob. 2, that the magnitudes of the electric and magnetic vectors are related by the same equation as for a plane wave, *viz.*,

$$\sqrt{\epsilon_0}\mathcal{E} = \sqrt{\mu_0}H$$

4. Starting with Eq. (11) of the text for the voltage distribution in a standing wave on a line of length $l/2$ (the fundamental mode), derive an expression for the current distribution using Eq. (6) of the text, and show that this is identical with the result expressed by Eq. (12) if $1/L'C' = c^2$.

5. A pair of parallel wires, each of length 3 meters, form a line which is short-circuited at *both* ends.

a. Compute the frequencies of the first three modes of oscillation of this system.

b. Write a general expression for the frequencies of the standing waves which may be set up on these wires.

6. A coaxial cable of length 4 meters, radius of the central conductor 2 mm. and inner radius of the outer cylinder 2 cm., is short-circuited at both ends. A standing wave is set up of frequency equal to that of the fundamental mode. The r.m.s. value of the current at one end of the cable (in the short-circuiting element) is 5.0 amp.

a. Write expressions for the energy density of the electric field and of the magnetic energy density at an arbitrary point inside the cable, and find an expression for the total energy density at this point.

b. Integrate your answer to part *a* to find the total electromagnetic energy of the wave, showing that it is constant, and compute this energy in joules.

7. An antenna of length l is driven at a frequency ν small compared to its lowest natural frequency, so that it carries an alternating current of amplitude I. Show that the expressions for the magnitudes of the electric and magnetic field intensities in the radiation field of this antenna take the form

$$|\mathcal{E}| = |H| = \frac{2\pi Il}{cr\lambda} \sin\theta \sin 2\pi\nu\left(t - \frac{r}{c}\right)$$

where Gaussian units are employed and λ is the wave length of the emitted waves.

Using the above equation derive the following formula for the average rate of energy radiation from the antenna:

$$\frac{\overline{dE}}{dt} \quad \frac{4\pi^2 I^2}{3c}\left(\frac{l}{\lambda}\right)$$

8. A radio station employs an antenna of length 30 meters and broadcasts on a wave length of 300 meters. If the antenna is to radiate 5 kw., compute the antenna current using the results of Prob. 7.

CHAPTER X

ELECTRONIC CONDUCTION IN VACUUM AND IN METALS

We have now reached the point in our study where we must turn to a discussion of the electrical and magnetic behavior of matter. Our principal task thus far has been the formulation of the fundamental laws of electromagnetism for empty space, and to accomplish this it has been necessary to introduce a number of facts concerning the electrical properties of matter, *e.g.*, the distinction between conductors and insulators and Ohm's law. These have been kept to a minimum, however, and now we proceed to a more detailed investigation of the laws governing the electromagnetic behavior of material media. The situation is somewhat analogous to that in which we found ourselves in the study of mechanics when we had completed our formulation of Newton's laws and their application to particles and to rigid bodies. Upon entering the field of the mechanics of deformable bodies, we found ourselves forced to inquire into the nature of the internal forces which hold matter together and found that there were two distinct modes of approach to this problem. (1) There is the large-scale viewpoint, treating material bodies as continuous media; and (2) there is the more fundamental, but more complicated, atomic viewpoint. Similarly, one can approach the problems of the electrical and magnetic behavior of matter in the same dual manner, and we shall do so in the following, but we shall spend more time on the atomic interpretation than we did in mechanics.

There is overwhelming evidence that atoms are composed of electrically charged particles, a central nucleus carrying practically all the mass of the atom and having a positive charge equal to an integral multiple of the charge on an electron, let us say Ze. Z is known as the atomic number, and it is this number which characterizes the chemical elements. Thus the hydrogen nucleus, the so-called *proton*, has $Z = 1$, the helium nucleus has an atomic number $Z = 2$, and so on throughout the whole

periodic table of the elements. Each atom possesses Z electrons, so that it is normally uncharged as a whole, and these are distributed more or less spherically around the central nucleus. The mass of the proton is about 1,840 times that of an electron. Atoms possessing a larger or smaller number of electrons than their normal complement of electrons are called *ions*, positive ions if they possess fewer than Z electrons and negative ions if they possess more than Z electrons. One essential point must always be kept in mind in thinking of any model of an atom, and that is that it is an open structure, the distances between the nucleus and the electrons and between the electrons being very large compared to the dimensions of either the nucleus or the electrons.

From an atomic standpoint the conduction of electricity in material bodies is due to motion of either electrons or ions or both. In electrolytes the current is due to the migration of both the positive and negative ions of the solute under the influence of the externally applied electric field. Migration of ions is also possible under the influence of strong fields in certain types of crystals in which there is a definite space lattice arrangement of positive and negative ions, as in the case of silver bromide. We shall concern ourselves in this chapter only with *electronic* currents. The simplest case occurs in the conduction of electricity by free electrons in high vacua such as one has in the ordinary radio tube. Somewhat more complicated is the conduction of electricity in metals, where electrons can move through the metal. These electrons, the so-called *free* or *conduction* electrons, can be thought of as the valence electrons of the metallic atoms which have been liberated from their parent atoms by the mutual interactions of the atoms when the latter are packed together as tightly as they are in a metallic lattice. These liberated electrons form a sort of gas and are more or less free, belonging to the metal as a whole rather than to individual atoms. Finally, in the case of electric discharges in gases, both mobile ions and electrons are present; the latter are mainly responsible for the current, but the presence of the former gives rise to extremely complicated phenomena.

47. Thermionic Emission; Electronic Currents in High Vacua. Before formulating the laws governing the flow of electronic currents in vacuum, we must say a few words concerning the

methods by which electrons can be liberated from a metal. In removing an electron from a metal, work must be done against the attractive forces which normally hold the electrons in the metal. The work per unit charge necessary to remove an electron from a metal is called the *work function* of the metallic surface and is of the order of magnitude of a few volts. In any case this work must be done if electrons are to be liberated and the various methods of liberation differ in the manner in which this energy is supplied. In the photoelectric effect the electrons gain energy from the absorption of light. Electrons liberated when a metallic surface is bombarded with electrons (so-called secondary electrons) pick up their energy from the impinging electrons. Positive-ion bombardment or bombardment by neutral atoms carrying more than their normal amounts of energy (metastable atoms) can also liberate electrons. Very intense electric fields can pull electrons out of metals (so-called cold emission), and finally heating the metal can impart enough thermal energy to some of the electrons so that they can escape in a manner analogous to the thermal evaporation of a liquid, and this phenomenon of *thermionic* emission has become familiar to all through the applications to radio tubes, etc.

Suppose we maintain a piece of metal with a plane surface at a definite temperature. We picture the electrons inside a metal as forming a gas and shall treat them as free. These free electrons inside the metal possess kinetic energy, and it was formerly supposed that they behaved just like the atoms of an ideal gas in that the average kinetic energy of an electron was $\frac{3}{2}kT$ (k is Boltzmann's constant) in accordance with the law of equipartition of energy. However, the fact that the contribution to the specific heat of a metal by its conduction electrons is exceedingly small always was a grave difficulty for the free-electron picture. We now know, however, that the electrons inside a metal do not follow the equipartition law and that their energy is practically independent of temperature, increasing but slightly when the metal temperature is raised. This small increase of energy, however, is just what is needed to enable the electrons to escape. Returning to our metallic sample, we then have the picture of some of the electrons escaping from the surface and forming a sort of a negatively charged cloud just outside the surface (Fig. 115). This negative *space charge*

inhibits the further release of electrons, and equilibrium is established. The density of electrons in the space outside the metal is so small compared to that inside the metal that these electrons do behave like an ideal gas and possess the normal amount of thermal energy, $\frac{3}{2}kT$ per electron. If these electrons are removed, by drawing them off to a collecting plate whose potential is maintained positive with respect to that of the emitting surface, more electrons escape, and there is a limiting rate (at each temperature) at which escape can take place. When the maximum electronic current flows from emitter to plate, we say that there is *saturation.* The saturation current density varies very markedly with temperature in close analogy with the evaporation

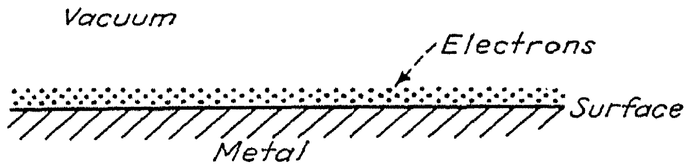

FIG. 115.

rate of a liquid. The law giving the saturation current density is known as *Richardson's equation,* and we present it without proof:

$$\frac{e\phi}{kT} \tag{1}$$

ϕ is the work function of the metal so that $e\phi$ is the gain of potential energy of an electron escaping from the metal. A is a constant depending on the metal but having a value of about 60 amp./cm.^{2}-°C.^{2} for many clean surfaces. Equation (1) has been tested over an enormous range of the variables involved; in fact it would be difficult to find any other electrical equation which has been verified over a wider range. The exponential dependence on temperature reminds one strongly of the vapor-pressure law for liquids.

When the collector (anode) potential is not high enough to cause the saturation current given by Eq. (1) to flow, a steady state is set up in which a smaller current flows, and one speaks of the current being limited by space charge. We shall investigate the laws governing the flow of space-charge limited currents for the simple case of plane parallel electrodes, the separation of which is small compared to the surface dimensions of either so that the electric field and potential vary only with an x-coordi-

nate normal to the planes of the electrodes as in the corresponding condenser problem. Since we are concerned only with a steady flow of current, we know that the electric field and potential obey the laws of electrostatics, and our first task is to formulate Gauss's theorem, which relates field intensity to charge density, in the form of a differential equation holding at any point of space. From this we can then obtain a relation between potential and charge density in the form of a differential equation known as *Poisson's* equation, which is more convenient for purposes of application than Gauss's theorem.

Poisson's Equation in One Dimension.—We apply Gauss's theorem in the form

$$\int D_n\,dS = 4\pi \int \rho\,dv$$

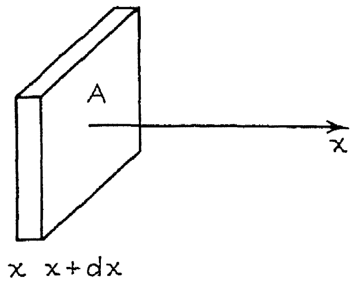

FIG. 116.

to a volume element shown in Fig. 116, for which the x-coordinate is normal to the faces A. The element has a thickness dx and is located at a position x. In our case D has only an x-component, and this can depend only on x. The flux of D emerging from the volume element shown in Fig. 116 is given by

$$\int D_n\,dS = (DA)_{x+dx} - (DA)_x = \frac{dD}{dx}A\,dx = \frac{dD}{dx}\,dv$$

so that Gauss's theorem requires that

$$\frac{dD}{dx} = 4\pi\rho$$

or

$$\frac{d\mathcal{E}}{dx} = \frac{4\pi\rho}{\epsilon_0} \tag{2}$$

using the fact that $D = \epsilon_0\mathcal{E}$.

Finally, since $\mathcal{E} = -\text{grad}\ V = -(dV/dx)$ for the case under consideration, Eq. (2) becomes

$$\frac{d^2V}{dx^2} = -\frac{4\pi\rho}{\epsilon_0} \tag{3}$$

which is Poisson's equation for one dimension. Note that if $\rho = 0$, denoting the absence of space charge, the integration of

Eq. (3) yields a uniform field in accordance with our former results for the parallel-plate condenser.

Returning to our original question of space-charge limited current, let the electrode separation be d, the hot cathode grounded (its potential equal to zero), and let the anode be maintained at a potential V_0 above ground (Fig. 117). The steady current is carried by a stream of electrons moving from cathode to anode, and the magnitude of the current density j is related to the velocity v of the electron stream and the charge density ρ by the equation

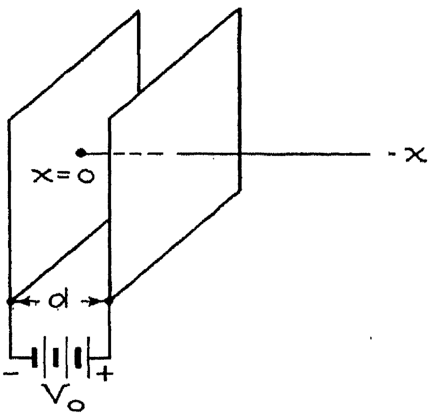

FIG. 117.

$$j = |\rho|v = nev \tag{4}$$

where n is the electron density at any point, v the velocity of the electrons at that point, and e is the magnitude of the electronic charge. In Eq. (4) both n and v vary from point to point, but their product is independent of position. If we denote by v_0 the average initial velocity of the electrons in the x-direction (as they leave the cathode), we have, according to the conservation of energy,

$$\tfrac{1}{2}mv_0^2 = \tfrac{1}{2}mv^2 - eV \tag{5}$$

or, solving for v,

$$v = \left(v_0^2 + \frac{2e}{m}V\right)^{\frac{1}{2}} \tag{6}$$

yielding a relation between electron velocity and potential at each point between the electrodes. Using the fact that the charge density ρ is related to the electron density n by $\rho = -ne$, Eq. (3) may be rewritten with the help of Eqs. (6) and (4) as

$$\frac{d^2V}{dx^2} = \frac{4\pi ne}{\epsilon_0} = \frac{4\pi}{\epsilon_0}j\left(v_0^2 + \frac{2e}{m}V\right)^{-\frac{1}{2}} \tag{7}$$

The variation of potential between the electrodes will thus be obtained if we can integrate this equation. Multiplying each side of Eq. (7) by $2(dV/dx)\,dx$ and making use of the identity

$$\frac{d}{dx}\left[\left(\frac{dV}{dx}\right)^2\right] = 2\frac{dV}{dx}\frac{d^2V}{dx^2}$$

we can integrate once and find readily

$$\left(\frac{dV}{dx}\right)^2 - \left(\frac{dV}{dx}\right)_0^2 = \frac{8\pi j}{\epsilon_0}\frac{m}{e}\left[\left(v_0^2 + \frac{2e}{m}V\right)^{\frac{1}{2}} - v_0\right] \tag{8}$$

where $-(dV/dx)_0$ is the field intensity at the cathode surface and we have set $V = 0$ for $x = 0$.

Thus far the equations which we have derived are quite general, holding equally well for the case of saturation current or for the case of space-charge limited current. What, then, is the physical distinction between these two cases, and how can we introduce these considerations into Eq. (8)? Let us start by considering Eq. (3). Since the right-hand side of the equation is always positive (ρ is negative because of the negative charge of the electron), the curve of potential against distance from the cathode must be concave upwards everywhere. This is equivalent to saying that the rate of change of the slope of the curve increases with increasing x. If the charge density ρ were negligibly small, V would be a linear function of x. If we imagine ρ to increase, the variation becomes something like that of curve (A) in Fig. 118, and, if ρ becomes large enough, a minimum will be present, as shown in curve (B). The critical curve which divides these two types is evidently the one for which the field and hence the slope becomes zero at $x = 0$. The potential distributions which have no minima correspond to saturation current, since the field is everywhere positive and every electron leaving the cathode will reach the anode. On the other hand, the presence of a potential minimum at P indicates a large electron density at this point (the curvature is greatest there), and the field near the cathode is reversed because of the repulsive force action of these electrons. In this case some of the electrons leaving the cathode (the slower ones whose initial energies are smaller than eV_{0p}) will turn around and only a fraction of the emitted electrons reach the anode. This, then, is the case of space-charge limited current. We now see that we must not only consider the average initial

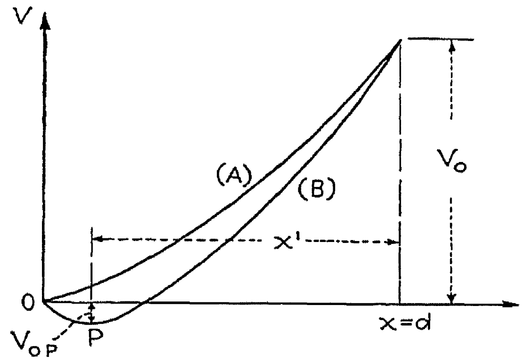

FIG. 118.

velocity of the electrons but that we must take into account the fact that the electrons leave the cathode with a distribution of velocities (in fact a Maxwell distribution) such that a larger and larger fraction return to the cathode the more pronounced the potential minimum. However, since the number of electrons leaving the cathode with initial energies greater than one electron-volt is entirely negligible for ordinary cathode temperatures, we need not concern ourselves with potential minima for which V_{0p} is greater than 1 volt.

It is then possible to obtain an approximate solution of Eq. (8) by shifting the origin of coordinates from O to P (Fig. 118). If the anode potential is of the order of 100 volts, we may consider V in Eq. (8) as measured from the potential at P without much error and also may neglect the initial velocity v_0 (now the velocity at P) compared to $\left(\frac{2e}{m}V\right)^{\frac{1}{2}}$. Since $\frac{dV}{dx}$ is zero at the point P, we set $(dV/dx)_0$ equal to zero in Eq. (8), and this equation becomes

$$\left(\frac{dV}{dx}\right)^2 = \frac{8\pi j}{\epsilon_0}\left(\frac{2m}{e}\right)^{\frac{1}{2}} V^{\frac{1}{2}}$$

or

$$\frac{dV}{dx} = \left(\frac{8\pi j}{\epsilon_0}\right)^{\frac{1}{2}}\left(\frac{2m}{e}\right)^{\frac{1}{4}} V^{\frac{1}{4}} \tag{9}$$

which yields upon integration

$$V = \left(\frac{9\pi j}{\epsilon_0}\right)^{\frac{2}{3}}\left(\frac{m}{2e}\right)^{\frac{1}{3}} x^{\frac{4}{3}} \tag{10}$$

satisfying the condition $V = 0$ at $x = 0$. Equation (10) gives very nearly the variation of potential between the point where the potential minimum exists and the anode. The current-voltage relation is obtained by noting that $V = V_0$ for $x = x'$. Inserting these values in Eq. (10) and solving for j, one obtains readily

$$j = \frac{\epsilon_0}{9\pi}\sqrt{\frac{2e}{m}}\,\frac{V_0^{\frac{3}{2}}}{x'^2} \tag{11}$$

Thus a current density proportional to the $\frac{3}{2}$ power of the potential difference between the electrodes is predicted, provided the value of x' does not vary appreciably with V_0. Actually the

potential minimum lies so close to the cathode that x' may be approximately replaced by d and Eq. (11) becomes

$$j = \frac{\epsilon_0}{9\pi}\sqrt{\frac{2e}{m}}\,\frac{V_0^{\frac{3}{2}}}{d^2} \tag{12}$$

and this is known as the Langmuir-Child equation. It can be shown that the proportionality between current and the $\frac{3}{2}$ power of the potential difference is not dependent on the geometrical arrangement of cathode and anode, subject to the limitations imposed by the approximations employed in our derivations.

48. Electrical Conductivity of Metals.—The electrical conductivity of metals is defined with the help of Ohm's law. In differential form this law can be written as

$$\vec{j} = \sigma\vec{\mathcal{E}} \tag{13}$$

In this section we shall show how our elementary picture of free electrons leads to Ohm's law and to an expression for the conductivity. In our picture we have free electrons wandering about among fixed positive ions with random velocities, very much as the molecules of a gas. On the average there is no resultant force on these conduction electrons and hence we imagine them moving in a region of constant potential, the average internal potential of the metal. If a uniform external field $\mathcal{E}$ is maintained inside the metal, there will be a force of magnitude $e\mathcal{E}$ acting on an electron. The electrons will be accelerated but will not move very far before colliding with the metal ions and losing the kinetic energy gained during this part of their motion. The average effect of these collisions with the ions is the same as if there were friction force acting on the electrons and we assume that this is like a viscous force, proportional to the electron speed. Thus the equation of motion of an electron becomes

$$e\mathcal{E} - kv = ma \tag{14}$$

Let us take the external field $\mathcal{E}$ in the x-direction as constant, and according to Eq. (14) the electron will drift with a constant velocity v given by

$$v \quad e\mathcal{E}$$

This is the steady-state solution of Eq. (14). If n is the number of electrons per unit volume, there results a steady current density

$$j = nev = \frac{ne^2}{k}\mathcal{E} \tag{15}$$

Comparing this with Eq. (13), we obtain for the conductivity

$$\sigma = \frac{ne^2}{k} \tag{16}$$

Thus the conductivity is proportional to the number of free electrons per unit volume. The above argument is exceedingly crude and neglects completely the kinetic energy possessed by the electrons in the absence of an external field. Actually this energy is very large compared with the amount gained by an electron between collisions with the ions. The average drift velocity of the electrons is very small compared with their random speed.

We can refine our calculation to take these facts into account and dispense with the rough picture of a viscous force. Let us assume that after each collision an electron, having given up the energy it has gained from the field to the lattice, starts again with its initial random speed. The time between two successive collisions will be different for different electrons, but we shall consider the average collision time which we denote by τ. This collision time is equal to l/u, where l is the mean free path of the electrons and u the average random speed. During the time τ an electron undergoes a uniform acceleration in the direction of the field, and the average velocity gained in excess of its initial component of velocity in this direction is

$$v = \frac{1}{2}a\tau = \frac{1}{2}\frac{e\mathcal{E}}{m}\tau = \frac{1}{2}\frac{e\mathcal{E}}{mu}l$$

This average excess speed in the direction of the applied field is just the drift velocity of the electrons, and the current density is

$$= nev = \frac{1}{2} \cdot \frac{\quad}{mu} \cdot l$$

Comparison with Ohm's law now gives for the conductivity

$$\frac{l}{2m\cdots} \tag{17}$$

This is a much more satisfactory formula than Eq. (16), and it is

essentially the one obtained by a much more elaborate theory based on modern statistical theory using a free-electron model. Since we know that the mean kinetic energy of the electrons in a metal is practically independent of temperature and the number of free electrons per unit volume probably does not vary appreciably with temperature, the actual temperature dependence of the conductivity must be explained by the variation of the mean free path l with temperature.

Actually, the conductivity of pure metals varies practically inversely with the absolute temperature, and there is good reason to believe that the mean free path varies this way also, at least at high enough temperatures so that the thermal vibrations of the metallic ions are practically independent of each other.

49. Thermoelectric Effects.—The most important thermal effect of an electric current flowing in a metallic conductor is the joule heating. This heating is an irreversible effect, independent of the direction of current flow. There are three very closely related *reversible* effects involving thermal and electrical energies which are denoted as the thermoelectric effects, and we shall examine them briefly in this section.

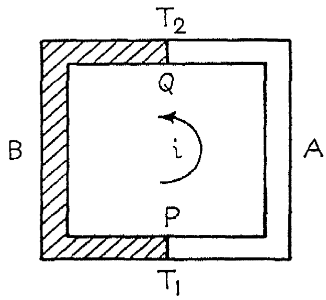

FIG. 119.

a. The Seebeck Effect.—If a closed circuit is constructed of two (or more) different metallic wires, a current will in general flow around the circuit if the junctions are maintained at different temperatures. This effect is termed the Seebeck effect after its discoverer. A simple circuit consisting of two metals A and B with junctions at P and Q is shown in Fig. 119, and let us suppose that the temperature T_2 of Q is higher than T_1, that of P. Such a device is known as a *thermocouple* and is used extensively as a temperature measuring instrument. The e.m.f. around this circuit is called a *thermal e.m.f.*, and it depends on the junction temperatures as well as on the two metals involved. We shall denote this thermal e.m.f. by E_{AB}, the order of the subscripts indicating that the e.m.f. is such that current flows from A to B *at the hot junction* Q. The interposition of a third metal C in series with the thermocouple circuit does not affect the thermal e.m.f. provided the junctions with this third metal are maintained at the same temperature.

b. The Peltier Effect.—When an electric current flows across a junction of two dissimilar metals, the temperature of the junction changes unless heat is supplied or abstracted by external means. The rate at which heat must be supplied to the junction to maintain constant temperature is proportional to the current and changes sign when the current changes direction. This phenomenon of the evolution or absorption of heat at junctions of dissimilar metals is called the Peltier effect. It occurs whether the current is driven across by an external agency or is spontaneously developed by the action of the thermocouple itself. Thus, in Fig. 119, the temperature T_2 of the junction Q tends to drop (heat must be supplied to maintain it), and the temperature T_1 of the junction P tends to rise. From the existence of this Peltier heat one infers the existence of an e.m.f. at the junction between two different metals. This *Peltier e.m.f.* will be denoted by Π_{AB}, indicating that the junction is between the metals A and B, and *a positive sign will mean that heat must be supplied to the junction to maintain its temperature when current flows from A to B.* This Peltier e.m.f. is in general a function of temperature. If a charge q is transported across the junction, the heat which must be supplied during this process is equal to Πq.

c. The Thomson Effect.—The third effect was predicted on the basis of theoretical thermodynamic arguments by Sir William Thomson (Lord Kelvin). If different parts of the *same* metallic conductor are maintained at different temperatures, a temperature gradient will exist inside the metal and a steady heat current will flow. If now an electric current flows in this metal, the temperature distribution will be disturbed, and the accompanying evolution or absorption of heat throughout the metal (in addition to the joule heating) comprises the Thomson effect. The Thomson heat which one must supply or abstract to maintain a steady temperature distribution is reversible, changing sign with change of direction of the electric current flow and it is proportional to the product of the temperature gradient and the electric current. *The existence of a temperature gradient inside a metal then implies the coexistence of an electric potential gradient.*

Consider an element of length dx of a wire, and let the temperature difference between the ends of this element be dT. If the *Thomson e.m.f.* due to this temperature gradient be denoted by dE', we define the *Thomson coefficient*, σ, by the relation

$$\sigma = \frac{dE'}{dT}$$

and call the coefficient *positive if the e.m.f. tends to drive current inside the metal along the temperature gradient, i.e., from low to high temperatures.* For a finite length of wire with its ends maintained at temperatures T_1 and T_2 ($T_2 > T_1$), the Thomson e.m.f. along the wire is evidently given by

$$E' = \int_{T_1}^{T_2} \sigma \, dT \tag{19}$$

The thermal e.m.f. of a thermocouple is the sum of the Peltier and Thomson e.m.fs. in the circuit, so that the Seebeck effect is not a phenomenon demanding separate explanation if one can explain the existence of the Peltier and Thomson effects. Consider the thermocouple circuit of Fig. 120. It is shown as a closed circuit for simplicity, but we can imagine it opened and connected to a potentiometer so that the thermal e.m.f. is measured without current flow. Suppose $T_2 > T_1$, so that the Peltier and Thomson e.m.fs. shown in the figure correspond to positive coefficients. From the figure it is evident that the net thermal e.m.f. E_{AB} is given by

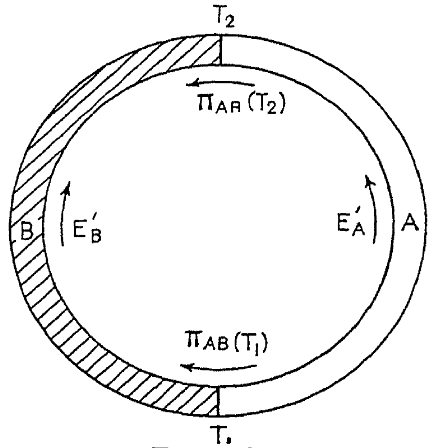

FIG. 120.

$$E_{AB} = \Pi_{AB}(T_2) - \Pi_{AB}(T_1) + E'_A - E'_B \tag{20}$$

or, using Eq. (19)

$$E_{AB} = \Pi_{AB}(T_2) - \Pi_{AB}(T_1) + \int_{T_1}^{T_2} \sigma_A \, dT - \int_{T_1}^{T_2} \sigma_B \, dT \tag{21}$$

It is often useful to write this relation in differential form. This may be obtained by applying the above relation to a thermocouple with its cold junction at temperature T and its hot junction at temperature $T + dT$. The result is evidently

$$e_{AB} = \frac{dE_{AB}}{dT} = \frac{d\Pi_{AB}}{dT} + (\sigma_A - \sigma_B) \tag{22}$$

where $e_{AB} = \dfrac{dE_{AB}}{dT}$ is called the *thermoelectric power* of the thermocouple. Equation (22) is essentially a statement of the

first law of thermodynamics as applied to a thermocouple. Application of the second law (we omit the derivation) yields

$$\frac{d}{dT}\left(\frac{\Pi_{AB}}{T}\right) + \frac{1}{T}(\sigma_A - \sigma_B) = 0 \tag{23}$$

where T is the absolute temperature. From Eqs. (22) and (23) follow the relations

$$\Pi_{AB} = T\frac{dE_{AB}}{dT} \tag{24}$$

$$\sigma_A - \sigma_B = -T\frac{d^2E_{AB}}{dT^2} \tag{25}$$

so that, from a knowledge of thermal e.m.fs. as functions of the temperature, one may compute the Peltier e.m.f. for the two metals from Eq. (24) and the difference of the Thomson coefficients from Eq. (25).

In tabulating the thermoelectric properties of metals, it is usual to use lead as a reference metal because its Thomson coefficient is zero. Suppose we have a thermocouple constructed of lead and some other metal and that one of the junctions is kept at 0°C. It turns out experimentally that the thermal e.m.f. of this thermocouple can be represented very closely by a quadratic function of the temperature of the hot junction. Thus we may write

$$E = at + \tfrac{1}{2}bt^2 \tag{26}$$

where t, the hot junction temperature, is expressed in degrees centigrade, and a and b are known as the Seebeck coefficients. The following table gives values for several metals referred to lead. A positive sign corresponds to current flow from lead to the metal at the hot junction so that lead corresponds to the metal A in our previous discussion.

Metal	a, microvolts/°C.	b, microvolts/°C.^{2}
Aluminum	− 0.47	+0.003
Bismuth	−43.7	−0.47
Copper	+ 2.76	+0.012
Gold	+ 2.90	+0.0093
Iron (soft)	+16.6	−0.030
Nickel	−19.1	−0.030
Platinum (Baker)	− 1.79	−0.035
Silver	+ 2.50	+0.012
Steel	+10.8	−0.016

In order to use the preceding table to compute the thermal e.m.f. of a thermocouple composed of two arbitrary metals A and B, one makes use of the following equation

$$= e_B - \tag{27}$$

where e_{AB} is the thermoelectric power of the thermocouple and e_A and e_B are the thermoelectric powers of thermocouples made of metal A and lead and of metal B and lead, respectively. The proof of this equation is left to the problems. It then follows that

$$e_{AB} = (a_B - a_A) + (b_B - b_A)t \tag{28}$$

where the a's and b's are the values listed in the foregoing table.

We must now inquire into an atomic interpretation of the Peltier and Thomson effects. Qualitatively, the free-electron gas picture offers a simple explanation of these effects. When a junction is made between two different metals (both at the same temperature), we are essentially bringing together two electron gases of different densities and pressures. There is a tendency for diffusion, and, since this does not occur in equilibrium, we must conclude that there is a difference of potential between two points lying on opposite sides of the boundary. This internal potential difference between two metals must then be equal in magnitude to the Peltier e.m.f. at the junction. In a crude way we may think of the Peltier e.m.f. as the equivalent of a seat of e.m.f. at the boundary, the nonelectrical forces acting in this seat of e.m.f. being diffusion forces. The work done by this seat of e.m.f. on the circuit when current flows is equal to the heat inflow to the junction from the surroundings.

As regards the Thomson effect, let us imagine that we have a straight rod of metal with one end in boiling water and the other in melting ice. A steady heat current will flow, and there will be a uniform temperature gradient in the metal. We think of the thermal conductivity of a metal being due to the transport of energy by the free electrons, so that we have the picture of the free electrons near the hot end of the metal gaining thermal energy in excess of those nearer the cold end. Now, since no electric current can flow, we must have a condition established in which equal numbers of electrons traverse a cross section of the rod per unit time in opposite directions. Those moving

toward the cold end have higher energies on the average than those moving in the opposite direction; hence there is a net transfer of energy, but not of charge, across this surface. In general, this balance will not be attained under the influence of the temperature gradient alone, but an electric field will be set up inside the bar of such magnitude and direction that the above conditions are satisfied. This field is to be identified with the Thomson potential gradient.

Problems

1. For clean tungsten the work function is 4.52 volts and the constant A in Richardson's equation is 60 amp./cm.²-°C.² Compute the saturation current density in amperes per square centimeter for tungsten at $T = 1000$°abs. and at $T = 2500$°abs. What is the ratio of these two currents?

2. For plane parallel electrodes of separation d, plot to scale the variation of potential and field as a function of distance from the cathode for the case of space-charge limited currents.

3. Space-charge limited current flows between two plane parallel electrodes, each of area 4 cm.² and separation 2.0 mm. Compute the total current when the potential difference between anode and cathode is maintained at 240 volts.

4. Assuming that the total space-charge limited current which flows between a hot cathode to an anode can depend only on e/m (ratio of charge to mass of an electron), on V_0 (the potential difference between anode and cathode), and on d (the separation of the electrodes), show by a dimensional analysis that the total current must be given by

$$i = B\epsilon_0 \left(\frac{e}{m}\right)^{\frac{1}{2}} V_0^{\frac{3}{2}}$$

where B is a dimensionless constant.

5. A copper wire of cross section 0.04 cm.² carries a steady current of 50 amp. Assuming that there is one free electron per atom of the metal, compute:

a. The number of free electrons per cubic centimeter.

b. The average drift velocity of these electrons.

c. The mean collision time of these electrons with the metallic ions.

The density of Cu = 8.9 grams/cm.³; the atomic weight of copper is 64, and its resistivity is 1.77×10^{-6} ohm-cm.

6. Derive Eqs. (24) and (25) of this chapter from Eqs. (22) and (23).

7. Prove that the thermoelectric power e_{12} of a pair of metals 1 and 2 is equal to the difference between the thermoelectric powers of these metals each taken in conjunction with the same reference metal [Eq. (27) of the text].

8. Show that for a thermocouple of two metals A and B, the thermal e.m.f. of the couple is given by

$$- t_1) + \frac{(b_B - b_A)}{2}(t_2^2 - t_1^2)$$

where t_1 and t_2 are the cold and hot junction temperatures, respectively, in degrees centigrade.

9. The "neutral" temperature of a thermocouple is the temperature at which the thermoelectric powers of the two metals (each with respect to lead) become equal, so that the thermal e.m.f. becomes a maximum with respect to varying temperature of the hot junction. Show that the thermal e.m.f. of the couple may be written in the form

$$E = e_0(t_2 - t_1)\left(1 - \frac{t_1 + t_2}{2t_n}\right)$$

where e_0 is the thermoelectric power of the thermocouple at 0°C. and t_n is the neutral temperature.

10. Compute the Peltier e.m.f. for a copper-nickel junction at 0°C. and the Thomson coefficients for each of these metals.

11. A gold-iron thermocouple has the following values for the thermoelectric powers for the gold and for the iron employed, respectively: $(2.8 + 0.01t)$ microvolts/°C. and $(17.5 - 0.048t)$ microvolts/°C. Calculate and plot the curve for the e.m.f. of this couple with one junction at 0°C. and the other at t°C., between 0°C. and 700°C. What is the temperature of the neutral point and the maximum e.m.f. of the couple?

12. A copper-iron thermocouple has a cold junction temperature of 20°C. Compute the maximum e.m.f. obtainable with this couple.

13. Prove that the thermal e.m.f. of a couple may be written as a quadratic function of the temperature difference between the junctions, *i.e.*, in the form

$$E_{AB} = \alpha(t_2 - t_1) + \frac{\beta}{2}(t_2 - t_1)^2$$

provided that

$$\alpha = (a_B - a_A) + (b_B - b_A)t_1$$
$$\beta = b_B - b_A$$

CHAPTER XI

DIELECTRICS

In this chapter we shall investigate the electrical behavior of insulating nonconducting material bodies, the so-called dielectrics. As previously pointed out, there is no sharp line of distinction to be drawn between conductors and insulators, but many substances, such as glass, waxes, and some crystals, are very nearly ideal insulators in that they can retain localized charges almost indefinitely. We shall assume in the following that we are dealing with ideal dielectrics and furthermore that they are both *isotropic*, the properties of the substance being the same in all directions at a point, and homogeneous. From an electronic viewpoint, dielectrics are characterized by the fact that the electrons are tightly bound to their parent atoms and cannot be dislodged with ordinary fields. As a result no conduction of electricity can take place by virtue of moving charges as in a conductor, and the conductivity of an ideal dielectric is taken to be zero. If an uncharged dielectric is located in a region of space where there is no electric field, the positive and negative charges in any small volume element produce no potential or field, and we may say that in such a volume element the center of "gravity" of positive and negative charges coincide. If, however, such a dielectric is placed in an external electric field, *e.g.*, between the plates of a charged condenser, then the positive charge is pushed in the direction of the field and the negative charge in the opposite direction. The forces holding the charges together may be thought of as elastic so that when such a separation of positive and negative charge takes place, there is a restoring force proportional to the separation distance, and a state of equilibrium is attained. Thus there is formed in each volume element a dipole (equal and opposite charges separated from each other), and this dipole is said to be *induced* by the external field, disappearing when the external field is removed. [The whole process of inducing dipoles in dielectrics is called

polarization. A polarized dielectric produces a field of its own, modifying the external field which gives rise to the polarization.

From an atomic standpoint the dipole moment induced in a volume element of a dielectric is thought of as the average of the dipole moments of all the atoms or molecules in that volume element. There are two cases which we must consider; (1) when the atoms or molecules possess no dipole moment in the absence of an external field and (2) the case of molecules which possess *permanent dipole moments* (the so-called *polar* molecules). For *nonpolar* molecules (the first case), the induced dipole moment per unit volume of the dielectric is simply the sum of the induced moments of the individual atoms or molecules. The induced dipole moment of an atom or molecule is proportional to the magnitude of the field producing it. The ratio of the induced dipole moment to the external field is called the *polarizability* of the atom or molecule. When molecules possess permanent dipole moments, the essential point is the question of the orientation of these dipole moments. In the absence of an external field the orientation is random; hence the average moment in a small volume element is zero, as indicated in Fig. 121*a*, where the arrows show the molecular dipole moments. When an external field $\mathcal{E}$ is applied, the dipoles line up in the direction of the field, as shown in Fig. 121*b*, since this is the position of stable equilibrium and the sum of the dipole moments is no longer zero. Actually the thermal motion of the molecules prevents them from lining up completely and tends to maintain the random orientation. The stronger the field $\mathcal{E}$ at a given temperature, the larger the dipole moment of the volume element and again we have a net result similar to the case of nonpolar molecules, the induced dipole moment increasing with the external field. For ordinary field strengths there is proportionality between the dipole moments produced and the field producing them. One can distinguish between the two types of polarization discussed above because one is strongly temperature dependent, the orientation effect decreasing with increasing temperature, whereas the other is independent of temperature. So much for the qualitative picture. We must now turn to a

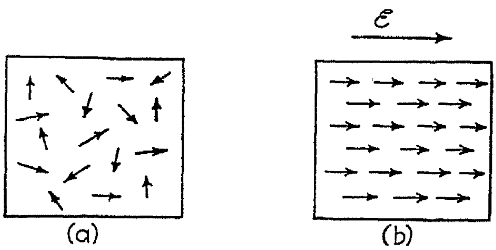

FIG. 121.

quantitative formulation of the laws governing the behavior of dielectrics.

50. Dielectric Constant; the Polarization Vector.—It was discovered by Cavendish, and later by Faraday, that, if a dielectric were inserted between the plates of a condenser, the capacity of the condenser increased. Let us imagine a charged condenser with insulated plates, and suppose we insert a dielectric between its plates. Since the charge on the plates $q = CV$ stays constant, the potential difference between the plates decreases, and this implies that the electric field inside the dielectric has become less than it was in vacuum. If the whole of the space between the condenser plates is filled with the dielectric, the capacity C' of the condenser becomes greater than its original capacity C by a factor κ, so that $C' = \kappa C$. κ is called the *dielectric constant* of the medium. Thus the potential difference and the field in the dielectric are reduced by the same factor. This result does not depend on the particular form of the condenser employed, and hence we infer that Coulomb's law for the force between point charges q and q' embedded in a dielectric has the form

$$F = \frac{qq'}{\kappa\epsilon_0 r^2} = \frac{qq'}{\epsilon r^2} \tag{1}$$

where $\epsilon = \kappa\epsilon_0$ is the permittivity or specific inductive capacity of the medium. In terms of the field intensity $\mathcal{E}$ due to a single point charge, we then have

$$\mathcal{E} = \frac{q}{\epsilon r^2} \tag{1a}$$

so that we can look upon ϵ as the ratio $D/\mathcal{E}$ [compare Chap. II, Eqs. (3) and (5)]. Note that κ is the dimensionless ratio ϵ/ϵ_0, and only in the electrostatic system of units is it identical with ϵ. κ is always greater than unity, only about 0.1 per cent for gases but is of the order of 2 to 10 for solid insulators, such as glass, and about 80 for pure water.

The reduction in electric intensity in a dielectric medium relative to its value in the absence of the medium is to be traced to the effect of the induced dipole field, which acts in addition to the external field. This field always acts in a direction opposite to that of the external field, and hence there results a lower value of the resultant electric intensity. This can be

readily seen with the help of a simple example. Suppose we have a slab of dielectric, as shown in Fig. 122, placed in a uniform external field $\mathcal{E}$. The fundamental fact is that, owing to the polarization of the medium, surface charges appear as indicated, equal and opposite on the two faces. This charge, although bound to the dielectric, contributes to the resultant field just as do other charges, such as those which we can place on conductors or move around as we wish. From the figure we see that the field of these induced surface charges opposes the external field $\mathcal{E}$.

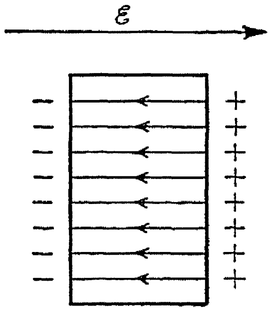

Fig. 122.

Let us now consider a medium in which we have equal and opposite charges e so close together that they produce no external effect. If an external field is applied, these charges are separated by a distance r and an induced dipole of moment

$$p = er \tag{2}$$

is created. We introduce a vector called the *polarization vector* P, which is defined as the *dipole moment per unit volume* induced in the medium. Clearly P is the vector sum of the induced atomic or molecular dipoles in a volume element divided by the volume of that element. Let us consider now an elementary area ΔA inside a dielectric, as shown in Fig. 123. If the medium is polarized by an external field acting in the direction indicated, charges will be pushed across ΔA and, in fact, the charge crossing ΔA will be equal to that originally contained in a slant prism of base ΔA and length r [the same r as in Eq. (2)]. If r_n is the component of r in the direction of the normal n to the surface element, then this charge is equal to

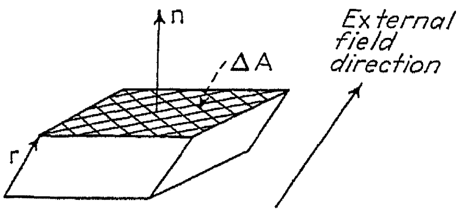

Fig. 123.

$$N(er_n)\,\Delta A = P_n \Delta A \tag{3}$$

where N is the number of dipoles per unit volume created in the dielectric. If ΔA should be an element of the outer surface of a body, then this represents the surface charge induced on the surface element. We thus have the important result that the surface density of charge σ_i induced on the surface of a polarized

dielectric is equal to the normal component of the polarization vector at that point. In symbols

$$\sigma_i = P_n \tag{4}$$

If we consider a closed volume element inside the medium, the net charge crossing the surface will be zero if the polarization is uniform (P = constant), but if the polarization varies from point to point, there may be more charge leaving than entering the volume element, with the result that one can have a charge density created inside this volume element, the so-called *polarization charge density*.

We can clarify the above statements with the help of a simple but important example. Consider a plane parallel condenser

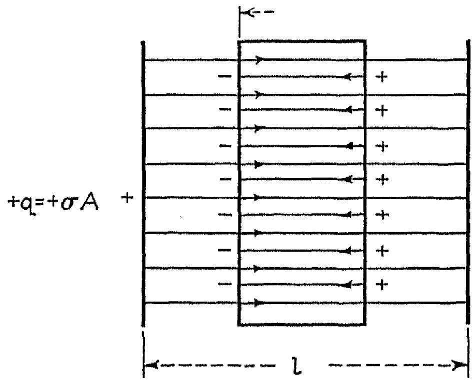

FIG. 124.

with a slab of dielectric material inserted between the plates, as shown in Fig. 124. Let the area of each plate and slab surface be A, the separation of the metal plates be l, and the slab thickness be d. In the figure are shown the lines of D starting on the "real" positive charges on the left-hand metallic plate and terminating on the negative charges on the right-hand plate. The lines of $\mathcal{E}$ inside the slab are to be obtained as the resultant of D (due to the "real" charges on the plates) and the field due to the induced surface charges $\pm\sigma_i A$ on the surfaces of the dielectric slab. Since the polarization vector is everywhere constant in magnitude and normal to the slab surfaces, we can write in accordance with Eq. (4),

$$\sigma_i = P$$

and hence the field intensity $\mathcal{E}$ inside the dielectric is given by

$$\mathcal{E} = \frac{4\pi}{\epsilon_0}(\sigma - \sigma_i) = \frac{4\pi}{\epsilon_0}(\sigma - P) \tag{5}$$

Since the slab faces are parallel to the condenser plates, the symmetry of the arrangement is not affected by the presence of the dielectric slab, with the result that the displacement vector D has the value $4\pi\sigma$ at every point between the plates, as is immediately evident from Gauss's law. Hence *in this example* the presence or absence of the dielectric slab does not affect the value of D. There then follows, from Eq. (5),

$$\mathcal{E} = \frac{D}{\epsilon_0} - \frac{4\pi P}{\epsilon_0} \tag{6}$$

or

$$D = \epsilon_0\mathcal{E} + 4\pi P \tag{7}$$

Equations (6) and (7) can be shown to hold in general, and they give the fundamental relation among the vectors $\mathcal{E}$, D, and P. Equation (6) in particular shows how the field $\mathcal{E}$ is reduced from its vacuum value D/ϵ_0 by the induced polarization field $4\pi P/\epsilon_0$ inside the dielectric.

Now let us suppose that the whole space between the condenser plates is filled with the dielectric, *i.e.*, that $d = l$. For this case, we have for the capacity of the condenser

$$C = \frac{\sigma A}{\mathcal{E} l} = \frac{DA}{\mathcal{E} 4\pi l} = \frac{\epsilon A}{4\pi l} \tag{8}$$

Since in vacuum this condenser would have a capacity $\epsilon_0 A/4\pi l$, the ratio of capacities is $\epsilon/\epsilon_0 = \kappa$, which checks the previous discussion. Note also that the total induced charge on the slab faces is $\pm PA$, and, since these faces are separated by a distance d, the dipole moment of these charges is PAd. Now Ad is simply the volume of the slab and we see for this case of uniform polarization that P is truly the induced dipole moment per unit volume.

We now have the following general picture: The field of D is that produced by charges distributed throughout space or put on conductors (controllable charge), the field of $\mathcal{E}$ is that produced by all charges, both of the controllable and polarization type, and the polarization field is due to the induced polarization charges; Eq. (7) giving the connection among the vectors $\mathcal{E}$,

D, and P. It is instructive to interpret Gauss's theorem for the flux of D with the help of these ideas. We have

$$\int_{\substack{\text{closed}\\\text{surface}}} D_n\, dS = 4\pi q \tag{9}$$

where q is the "real" controllable charge inside the closed surface. Using Eq. (7) for D, the above equation can be written as

$$\epsilon_0 \int_{\substack{\text{closed}\\\text{surface}}} \mathcal{E}_n\, dS + 4\pi \int_{\substack{\text{closed}\\\text{surface}}} P_n\, dS = 4\pi q \tag{10}$$

Now $\int P_n \cdot dS$ is just the charge which has crossed the closed surface due to polarization according to Eq. (3), and hence there remains a polarization charge inside the volume equal to

$$q_p = -\int P_n\, dS$$

as it must be of opposite sign to that which has passed out of the volume. Putting this into Eq. (10), we find for the flux of $\mathcal{E}$ emerging from the volume

$$\int_{\substack{\text{closed}\\\text{surface}}} \mathcal{E}_n\, dS = \frac{4\pi}{\epsilon_0}(q + q_p) \tag{11}$$

which checks our statement that $\mathcal{E}$ was the field of all the charge, "real" plus polarization.

If we write $\epsilon\mathcal{E}$ for D in Eq. (7), we have

$$D = \epsilon\mathcal{E} = \epsilon_0\mathcal{E} + 4\pi P$$

or

$$\epsilon = \epsilon_0\left(1 + \frac{4\pi P}{\epsilon_0\mathcal{E}}\right) \tag{12}$$

or again

$$\kappa = \frac{\epsilon}{\epsilon_0} = 1 + \frac{4\pi}{\epsilon_0}\frac{P}{\mathcal{E}} \tag{12a}$$

Thus the dielectric constant κ will truly be constant if the polarization vector P is proportional to the field intensity $\mathcal{E}$. The ratio $P/\epsilon_0\mathcal{E}$ is known as the *electric susceptibility* χ of the medium,

$$P = \epsilon_0\chi\mathcal{E} \tag{13}$$

so that the relation between dielectric constant and susceptibility is

$$\kappa = 1 + 4\pi\chi \tag{14}$$

51. Cavity Definitions of D and $\mathcal{E}$.—Let us now inquire somewhat more closely as to exactly what one means by the intensity of an electric field inside a dielectric medium. We recall that the intensity $\mathcal{E}$ for empty space was defined as a vector specifying the force per unit charge on an infinitesimal test charge placed at the point in question. In order properly to define $\mathcal{E}$ inside a dielectric, it becomes necessary to provide for the possibility of introducing a test charge. This is done, following Lord Kelvin, by imagining that one scoops out a cavity inside the dielectric. Then one can introduce a tiny charge in the cavity and determine the force on it, at least in principle. In Fig. 125 we show the parallel-plate condenser with the slab of dielectric. If we imagine a long needle-shaped cavity such as a with its long sides parallel to the field, then a test charge placed in it would be acted on by the field of both the real and polarization charges, and thus the force would be $q\mathcal{E}$ if q is the magnitude of the point test charge. It is necessary to choose the cavity of the form and orientation specified to avoid the appearance of polarization bound charge on the cavity surface. Thus in a cavity of the form b, which we imagine shaped like a pillbox with its flat faces perpendicular to the field, there will appear bound surface charges on these faces of density $\pm P$, so that the lines starting from the dielectric slab surfaces will terminate on the faces of the cavity and none will traverse it. As a result, only the real charge on the metal plates is effective in producing the force on a test charge in a cavity of shape b and this force will be

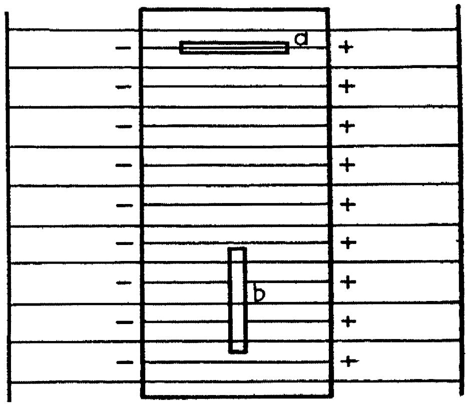

FIG. 125.

$$q_{\epsilon_0} \qquad + \frac{4\pi P}{\quad}\Big) = q\kappa\mathcal{E}$$

using Eq. (12a). The reason for the needle-shaped cavity a is

to minimize the effect of polarization charge on its end faces, which can be done by making them arbitrarily small in comparison with the length of the cavity. The field inside a cavity in a dielectric depends on the shape of the cavity; for example, in the case of a spherical cavity the field at the center may be shown to be

$$[\mathcal{E} + (4\pi/3\epsilon_0)P]$$

52. The Dielectric Constant of Gases.—There is one case for which it is not difficult to compute the dielectric constant on the basis of a simple atomic picture. This is the case of a gas, let us say a monatomic gas, at moderate pressures, so that the atoms are relatively far enough apart from each other on the average to allow us to neglect the interactions between them. Each atom contains charges $-Ze$ on the electrons, and $+Ze$ on the nucleus, which can be displaced relative to each other in an external field and behave as if held together by linear restoring forces. Suppose a uniform external field $\mathcal{E}$ is applied to the gas, a charge q is acted on by the force $q\mathcal{E}$ due to the field, and a restoring force $-kx$, if x is the direction of the uniform field $\mathcal{E}$. In equilibrium we have

$$-kx + q\mathcal{E} = 0$$

or

$$x = \frac{q\mathcal{E}}{k}$$

and the dipole moment created is

$$p = qx = \frac{q^2}{k}\mathcal{E} \tag{15}$$

The *polarizability* α of an atom is defined, as previously mentioned, by the ratio of the induced dipole moment to the field intensity, so that

$$\alpha = \frac{p}{\mathcal{E}} = \frac{q^2}{k} \tag{16}$$

If there are n atoms per unit volume, the polarization vector P is simply $np = n\alpha\mathcal{E}$, and we have from Eq. (12a)

$$\kappa = 1 + \frac{4\pi}{\epsilon_0}n\alpha \tag{17}$$

Now let us try to get an idea of the order of magnitude of the atomic polarizability α. To do this we shall adopt a simple

model of the atom, a central point charge $+Ze$ corresponding to the nucleus, surrounded by a uniform spherical distribution of radius R of negative charge $-Ze$ corresponding to the electrons. Under the influence of an external field the positive charge will move a distance x in the direction of the field relative to the negative charge, from O to P in Fig. 126. When in this position, it will be attracted back toward O by the negative charge. From our previous work in electrostatics we can calculate this force by considering all the negative charge inside the dotted sphere of radius x to be concentrated at O, the negative charge outside this sphere yielding no contribution to the force. The negative charge at the center O is then

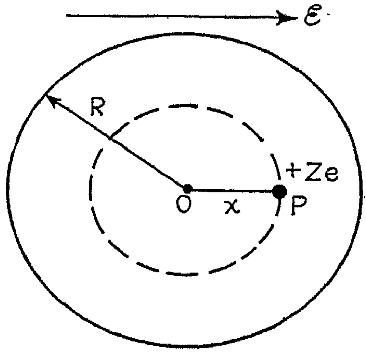

Fig. 126.

$$q' = -$$ [illegible]

and from Coulomb's law the attraction force is

$$F = -\frac{Z^2e^2}{\epsilon_0 x^2}\left(\frac{x^3}{R^3}\right) = -\frac{Z^2e^2}{\epsilon_0 R^3}\cdot x = -kx$$

and this is a linear restoring force of the type we used in the foregoing discussion. From Eq. (16) we then find for the polarizability, since $q = Ze$,

$$\alpha = \frac{(Ze)^2}{k} = \epsilon_0 R^3 \tag{18}$$

and it is simply proportional to the volume of the atom. Using this value of α in Eq. (17), we find the dielectric constant of a gas to be given on the basis of this approximate model by

$$\kappa = 1 + 4\pi n R^3 \tag{19}$$

The number of atoms per cubic centimeter of a gas is about 3×10^{19} under standard conditions, and for the atomic radius R we may use a rough value of 10^{-8} cm. = 1 angstrom. Thus the term $4\pi nR^3$ becomes approximately 0.4×10^{-3} or about 0.04 per cent. This is the actual order of magnitude of $\kappa - 1$ for

53. Boundary Conditions on D and $\mathcal{E}$.—Our examination of the manner in which the electric intensity and displacement

vectors change as one moves across a boundary separating two dielectric media has thus far been confined to the special case in which the interfacial surface is normal to the direction of these field vectors. We must now examine what happens when the field vectors are no longer normal to the surface of separation and in so doing we shall be able to understand why the law of Coulomb is not valid under all conditions. Thus the law of Gauss becomes one of the fundamental and universally applicable laws of electricity. Let us consider the case of two dielectric media of different permittivities, ϵ_1 and ϵ_2, and apply Gauss's theorem to the infinitesimal volume enclosed by the pillbox shown in Fig. 127. The pillbox is so chosen that a portion of the boundary surface lies inside the volume and we can, by making the pillbox infinitely shallow, neglect the flux of D across its curved surface, which is normal to the boundary between the dielectrics. For the two flat faces, we have for the flux of D emerging from the volume

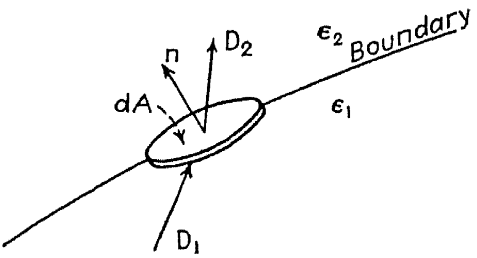

FIG. 127.

$$d\psi = (D_{2n} - D_{1n})\,dA$$

so that Gauss's theorem yields

$$(D_{2n} - D_{1n})\,dA = 4\pi\sigma\,dA$$

where σ is the surface density of charge on the interface. D_{1n} and D_{2n} are the components of D normal to the boundary surface at the point in question. Thus we have

$$D_{2n} - D_{1n} = 4\pi\sigma \tag{20}$$

or, if there is no surface charge on the interface,

$$D_{1n} = D_{2n} \tag{21}$$

In words, the lines of the electric displacement vector must have continuous normal components when crossing an uncharged boundary surface between two dielectrics. Applied to the example of the dielectric slab between the plates of the condenser (Fig. 124), this condition states that the value of D is unchanged as one crosses the surface of the slab, since the lines of D are normal to this surface. Evidently, this will be true

in all cases for which the boundaries are normal to the direction of the field.

For the components of the field tangent to the boundary, we must apply the Faraday induction law to a closed path such as is shown in Fig. 128, in which the sides perpendicular to the boundary are made infinitely short compared to the parallel sides dl. As the lengths of the short sides are made smaller and smaller, the area enclosed by the path approaches zero, so that the magnetic flux through this area also approaches zero. Hence

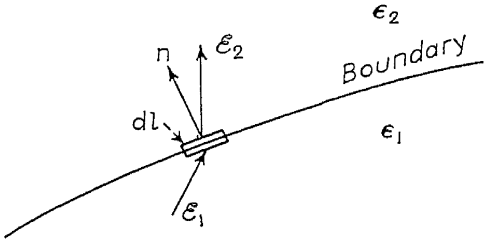

FIG. 128.

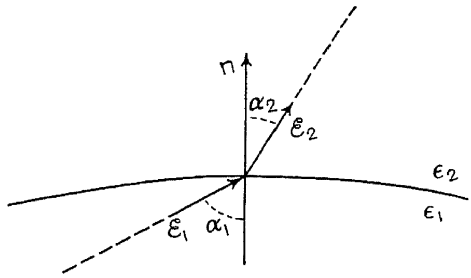

FIG. 129.

the Faraday law requires that the e.m.f. around this path be zero. We then may write

$$\oint \mathcal{E}_s \, ds = \mathcal{E}_{1t} \, dl - \mathcal{E}_{2t} \, dl = 0$$

or

$$\mathcal{E}_{1t} = \mathcal{E}_{2t} \tag{22}$$

where $\mathcal{E}_{1t}$ and $\mathcal{E}_{2t}$ are the tangential components of the electric intensity in media 1 and 2 at the point in question. Note that this condition is automatically satisfied in cases for which the boundary is normal to the field, such as the example of Fig. 124.

The two boundary conditions (21) and (22) are sufficient to determine uniquely the relative directions of the lines of $\mathcal{E}$ or D as one passes across an uncharged boundary between two dielectrics in any case. Let α_1 and α_2 be the angles between the normal to the boundary and the directions of $\mathcal{E}_1$ and $\mathcal{E}_2$ at the point where the normal is drawn as shown in Fig. 129. We write, in accordance with Eq. (22),

$$\mathcal{E}_1 \sin \alpha_1 = \mathcal{E}_2 \sin \alpha_2 \tag{23}$$

and, according to Eq. (21), using $D = \epsilon\mathcal{E}$,

$$\epsilon_1\mathcal{E}_1 \cos \alpha_1 = \epsilon_2\mathcal{E}_2 \cos \alpha_2 \tag{24}$$

Dividing Eq. (23) by Eq. (24), there follows

$$\frac{\tan \alpha_1}{\tan \alpha_2} = \frac{\epsilon_1}{\epsilon_2} = \frac{\kappa_1}{\kappa_2} \tag{25}$$

and this is the law of refraction for the electric field at the boundary between two dielectrics of dielectric constants κ_1 and κ_2, respectively. It is obvious that the same law gives the relative direction of the lines of D. If the direction of the field at a boundary is known in one medium, Eq. (25) uniquely fixes its direction at that boundary in the second medium. Note

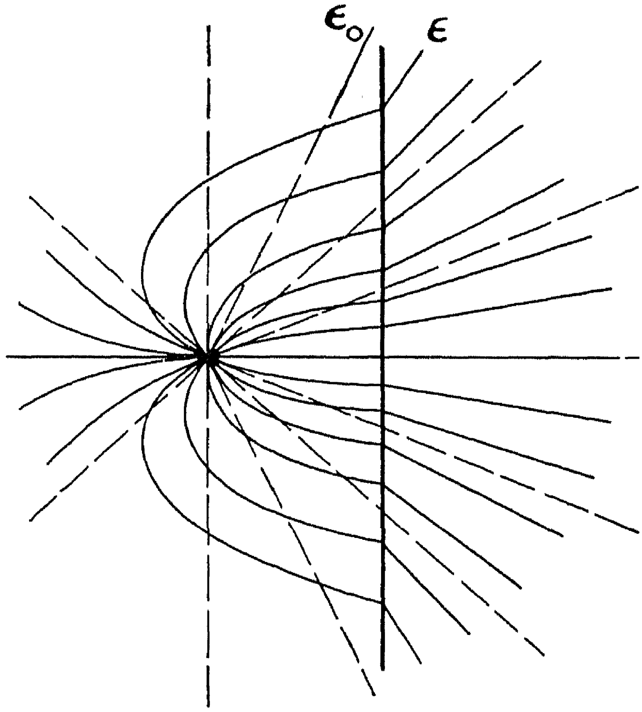

FIG. 130.

that the lines of force are bent more toward the normal in the medium of smaller dielectric constant than in the other medium. Thus Fig. 129 is drawn for the case $\kappa_1 > \kappa_2$.

We are now in a position to state precisely the limitations which must be imposed on the simple definition of D as given by Eq. (3) or Eq. (4) of Chap. II (Coulomb's law) and on the statement that the field of D is determined solely by the positions and magnitudes of the "real" charges. These statements are true only (1) if the field produced by the charges is appreciably different from zero in empty space or in a single homogeneous dielectric medium or (2) if the boundaries of the dielectrics

present are everywhere normal to the direction of the field which would be set up in the absence of these dielectrics. In the presence of arbitrarily shaped dielectrics of finite size, the only property of the field of D which is unaltered by these dielectrics is the total flux of D, due to the "real" charges present (Gauss's law). At the boundaries the direction of D undergoes a change in general, but the number of lines of D is unaltered by the presence of the boundary. We thus see why Gauss's law is so important. The above statements may be clarified with the help of a simple example. Consider the field produced by a single point charge q in the presence of an infinite slab of dielectric as shown in Fig. 130.

In the absence of the dielectric, the lines of D are radial and are shown dotted in the figure. The solid lines show the actual field of D in the presence of the dielectric. One sees that the direction (and magnitude) of D is altered even in empty space by the presence of the dielectric. However, the total flux of D is $4\pi q$ in any case.

The solution of the general problem, *e.g.*, one such as that shown in Fig. 130, involves the simultaneous solution of the equations

$$\int_{\substack{\text{closed}\\\text{surface}}} D_n\,dS =$$

$$\oint \mathcal{E}_s\,ds = 0$$

$$D = \epsilon\mathcal{E}$$

for the case of electrostatic fields, subject to the boundary conditions derived in this section. This would take us far beyond the scope of this book.

54. Polarization and Displacement Current in Dielectrics.— We now turn to a brief discussion of transient or nonsteady fields in dielectric media. For simplicity let us first consider the case of a parallel-plate condenser with a dielectric of permittivity ϵ filling the space between the plates, as shown in Fig. 131. Suppose the switch S is closed and a charging current starts flowing. As the field between the condenser plates builds up, the dielectric becomes polarized, positive charge moves, as shown, across the surface aa, and negative charge moves in the opposite direction. This motion corresponds to a transient

current while the dipoles are being formed, and this current is called polarization current. The polarization current can be expressed in terms of the polarization vector P. From Eq. (3) it follows that the charge crossing the area $aa = A$ in time dt is $A\ dP$, where dP is the increase in the polarization vector in this time. Consequently the polarization current is

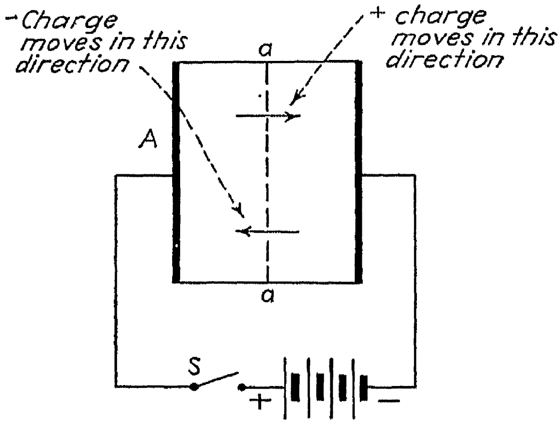

FIG. 131.

$$i_p = A\frac{\partial P}{\partial t} \tag{26}$$

and the corresponding current density is

$$j_p = \frac{\partial P}{\partial t} \tag{27}$$

This polarization current contributes to the total displacement current flowing between the condenser plates. The displacement current is given by $(A/4\pi)(\partial D/\partial t)$ and, using Eq. (7), can be written as

$$i_d = \frac{\epsilon_0 A}{4\pi}\frac{\partial \mathcal{E}}{\partial t} + A\frac{\partial P}{\partial t} \tag{28}$$

Thus the displacement current can be thought of as the sum of two terms, one a displacement current $\frac{\epsilon_0}{4\pi}A\left(\frac{\partial \mathcal{E}}{\partial t}\right)$ which would flow in vacuum and the other the polarization current in the dielectric. The ratio of the polarization to the vacuum displacement current is then, using Eqs. (13) and (14),

$$\frac{A(\partial P/\partial t)}{(\epsilon_0 A/4\pi)(\partial \mathcal{E}/\partial t)} = \frac{4\pi\epsilon_0 A\chi(\partial \mathcal{E}/\partial t)}{\epsilon_0 A(\partial \mathcal{E}/\partial t)} = 4\pi\chi = \kappa - 1 \tag{29}$$

Thus in the case of the medium of dielectric constant $\kappa = 2$, one-half the displacement current is polarization current.

More generally, if we consider the case of nonuniform polarization of a dielectric, the total polarization current flowing across a fixed closed surface is given by

$$i_p = \frac{\partial}{\partial t}\int_{\substack{\text{closed}\\ \text{surface}}} P_n\, dS \tag{30}$$

and we have seen from the considerations leading from Eq. (10) to Eq. (11) that

$$\int P_n\, dS = -q_p$$

where q_p is the polarization charge inside the closed surface. Consequently we may write

$$_n\, dS = -\frac{\partial q_p}{\partial t} = -\frac{\partial}{\partial t}\int \rho_p\, dv \tag{31}$$

where ρ_p is the density of polarization charge. Equation (31) is just the equation of continuity for polarization charge and current, stating that the current flowing out of the volume equals the rate of decrease of charge inside this volume.

Problems

1. Consider a parallel-plate condenser with a medium of permittivity ϵ between its plates. Following the arguments of Sec. 16, Chap. III, show that the electrostatic energy density in the medium is given by

$$u = \frac{\quad}{8\pi} \qquad \frac{\quad}{8\pi\epsilon}$$

and that the force of attraction of one plate for the other is given by

$$F = \frac{2\pi A\sigma^2}{\epsilon} = \frac{\epsilon\mathcal{E}^2}{8\pi}A$$

2. A parallel-plate air condenser is connected to a battery which maintains a difference of potential V_0 between its plates. A slab of dielectric of dielectric constant κ is inserted between the plates, completely filling the space between them.

a. Show that the battery does an amount of work $V_0\, q_0(\kappa - 1)$, if q_0 is the charge on the condenser plates before the slab is inserted.

b. How much work is done by mechanical forces on the slab when it is inserted between the plates? Is this work done on, or by, the agent inserting the slab?

3. A parallel-plate condenser with plate areas of 200 cm.² and a separation of 2.0 mm. is immersed in oil of dielectric constant 3.0 and permanently connected to a 300-volt battery.

a. Compute the charge on the condenser plates.

b. Compute the induced dipole moment per unit volume and the electric field intensity in the oil between the plates. State units.

c. What is the force of attraction of the plates for each other?

d. If the plates are separated to a distance of 4.0 mm., maintaining the potential difference constant at 300 volts, calculate the mechanical work done in effecting this separation.

e. How much energy is supplied to the condenser by the battery during the process described in part *d*?

4. A parallel-plate condenser of separation d has a capacity C_0 in air. A slab of dielectric of dielectric constant κ, thickness $t < d$, and area equal to that of the plates is introduced between the plates, the faces of the slab being parallel to those of the condenser. Neglecting end effects, prove that the capacity of the condenser becomes

$$C = \frac{C_0}{1 - \frac{t}{d}\left(\frac{\kappa - 1}{\kappa}\right)}$$

5. A condenser is formed of two concentric spherical metal shells of radii 2 and 6 cm. The inner sphere is covered by a wax coating 3 cm. thick, and the remainder of the space between the spheres is filled with a liquid of dielectric constant 4.2. The dielectric constant of the wax is 2.0.

Compute the capacity of the condenser thus formed in microfarads.

6. If the plates of the condenser of Prob. 5 are maintained at a potential difference of 3,000 volts, compute the total energy stored in the condenser. What is the surface density of polarization charge at the wax-liquid interface?

7. The inner sphere of a spherical condenser of inner and outer radii a and b, respectively, is coated with a thin coat of varnish of thickness t and dielectric constant κ. Show that the increase of capacity due to the varnish is approximately given by

$$\epsilon_0\left[\frac{b^2(\kappa - 1)}{\kappa(b - a)^2}\right]t$$

8. The dielectric constant of the material between the plates of a parallel-plate condenser varies uniformly from one plate to the other. If κ_1 and κ_2 are its values at the two plates, prove that the condenser has a capacity

$$C = \frac{\epsilon_0 A}{4\pi d} \cdot \frac{\kappa_2 - \kappa_1}{\ln(\kappa_2/\kappa_1)}$$

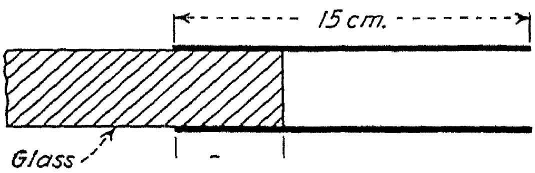

FIG. 132.

9. A parallel-plate condenser consists of two square plates 15 cm. on a side separated by a distance of 0.3 cm. A slab of glass of dielectric constant 6.0, thickness 0.3 cm. and 15 cm. on a side, is inserted, as shown in Fig. 132, between the condenser plates, and the condenser is connected to a 600-volt battery. Neglect the edge effects.

a. Compute the capacity of the condenser so formed.

b. What is the total charge on the condenser plates, and how is it distributed?

c. How much energy is stored in the condenser?

d. Suppose the battery is disconnected, leaving the condenser charged, and the glass slab moves in a distance dx farther to the right. Obtain an expression for the increase or decrease of energy in the condenser.

e. From your answer to part *d* compute the force tending to pull the glass slab between the plates. How does it depend on the length of slab already inserted?

10. Suppose that the thickness of the glass slab in Prob. 9 is 0.2 cm. and that the battery is left connected to the plates.

a. Compute the force tending to draw the slab between the plates.

b. How much work is done on the slab if it starts as shown in the figure of Prob. 9 and ends with 10 cm. of its length between the plates?

c. How much energy does the battery supply to the system during this process?

d. Compute the change of field energy in the condenser during the process described in part *b*, and show that it is equal to the energy supplied by the battery minus the mechanical work done on the glass slab.

11. A parallel-plate condenser of plate separation d in air is charged by a battery, and the battery is then disconnected. A slab of dielectric of thickness $t < d$ and area equal to that of either of the condenser plates is introduced between the plates, the slab faces being parallel to the condenser plates.

Prove that the electrostatic energy in the condenser field is *decreased* by an amount given by

$$\frac{v}{2}P\mathcal{E}_0$$

where v is the volume of the dielectric, P the polarization vector in the dielectric, and $\mathcal{E}_0$ the field intensity which existed before the insertion of the dielectric.

12. Consider the electrostatic field set up in air by a number of fixed charged conductors. If a small rigid dielectric body is introduced into the field in a position far enough from the conductors so that the distribution of true charge on these conductors is not sensibly altered by the introduction of the dielectric, the field energy may be shown in general to change by an amount

$$U - U_0 = -\tfrac{1}{2}\int P\mathcal{E}_0\, dv$$

(the result of Prob. 11 is a special case of this general formula). Given the fact that the field intensity $\mathcal{E}$ inside a dielectric sphere is uniform when the sphere is placed in an originally uniform field of intensity $\mathcal{E}_0$, and that $\mathcal{E}$ is related to $\mathcal{E}_0$ by the equation

$$\mathcal{E} = \frac{3}{\kappa + 2}\mathcal{E}_0$$

derive an expression for the change of field energy caused by the introduction of a dielectric sphere of radius a into a position where the field originally was $\mathcal{E}_0$. Take the volume of the sphere small enough so that $\mathcal{E}_0$ does not sensibly vary over the region of space occupied by this volume.

13. Let the dielectric sphere of Prob. 12 be displaced slightly to a point where the original field had a value slightly different from $\mathcal{E}_0$. Compute

the change in electrostatic energy, and, using the fact that this must equal the mechanical work done on the sphere, show that the mechanical force F acting on the sphere is given by

$$\vec{F} = \frac{\epsilon_0 a^3}{2} \frac{(\kappa - 1)}{(\kappa + 2)} \operatorname{grad} \mathcal{E}_0^2$$

14. A small dielectric sphere of susceptibility χ and radius b is placed at a large distance r from a metal sphere of radius a which is maintained at a potential V. Assume that $b \ll r$ and that $a \ll r$. Using the result of Prob. 13, derive an expression for the force with which the metal sphere attracts the dielectric sphere.

15. The vertical plates of a parallel-plate condenser are dipped into an insulating liquid of susceptibility χ. Neglecting surface tension, show that, if a potential difference V is established between the plates, the liquid will rise between the plates to a height h above its initial level where h is given by

$$h = \frac{\chi \epsilon_0 V^2}{2\rho g d^2}$$

d is the separation of the plates and g is the acceleration of a freely falling body. The susceptibility of air is neglected.

16. Prove that, at an interface between two dielectrics of dielectric constants κ_1 and κ_2 which carries no true charge, there is a surface polarization charge density σ_p given by

$$\sigma_p = \frac{\epsilon_0}{4\pi} \mathcal{E}_{1n} \left(1 - \frac{\kappa_1}{\kappa_2} \right)$$

where $\mathcal{E}_{1n}$ is the component of $\mathcal{E}$ in medium 1 normal to the interface at the point in question.

CHAPTER XII

MAGNETIC MEDIA

The study of magnetism was pursued as a branch of physics entirely distinct from that of electricity up to the time of the discoveries of Oersted and Faraday and probably is the older of the two subjects. The early study of magnetic fields concerned itself with the interactions of permanent magnets and particularly with terrestrial magnetism. The laws governing the behavior of the magnetic fields of permanent magnets were formulated in a manner analogous to the laws of electrostatics, at least as far as was possible, and even today many expositions of the subject are treated on the basis of these analogies. In our study, however, we have introduced the magnetic field vectors, at least for empty space, in terms of the electric currents which give rise to them, and we shall continue to adopt this mode of interpretation even for the case of magnetized material bodies. The formulas which we shall find will be essentially the same as those obtained by the more classical treatment, but the interpretation we give them will be based on our present-day atomic ideas concerning the origin of the magnetic behavior of material media. This mode of interpretation is not really new, since Ampère pointed out the possibility of utilizing it. In Ampère's time, however, and until comparatively recently, there was no particular reason to adopt one mode of interpretation rather than the other, but today we have a large amount of evidence showing that the electrons in matter are responsible for its observed magnetic behavior.

Returning to the question of the analogy between electric and magnetic fields, the starting point of electrostatics is Coulomb's law with the consequent possibility of defining and obtaining units of charge and electric intensity. Right at the start the analogy between electrostatics and magnetic fields breaks down completely, since it is impossible to produce a "magnetically charged" body. (This has not hindered writers from assuming its existence, however.) It is true that one

can magnetically polarize a material body so that the external magnetizing field is modified by its presence, but one cannot impart a "magnetic charge" to such a body. On the other hand, there are substances, called magnetically hard, which can be magnetized and retain some of the induced magnetization when the external field is removed. These substances, permanent magnets, now produce magnetic fields of their own, and their permanent magnetization is almost independent of external fields, at least for weak fields. We can imagine ideal magnetically hard permanent magnets which are not affected at all by external fields and then have the possibility in principle of utilizing a tiny permanent bar magnet to investigate and define magnetic field strength and magnetic moments. In so doing, the test bar is treated both as a "source" (a dipole source) and as an indicator of the magnetic field.

In Sec. 31 of Chap. V we have introduced the idea of the magnetic moment of a tiny current loop or "whirl" and can carry this discussion over as it stands to examine the behavior of our tiny permanent magnet test bar, thus indicating from the start a possible interpretation of the magnetization of material bodies. If a test bar is suspended by a thread fastened to its center, it will align itself with the direction of the earth's magnetic field at the point where it is located and, if displaced from equilibrium, oscillates with a frequency

$$n = \frac{1}{2\pi}\sqrt{\frac{mB_0}{I}} \tag{1}$$

[see Eq. (52), Chap. V], where m is the magnetic moment of the bar, I its moment of inertia about the axis of suspension, and B_0 the magnetic induction vector of the earth's field. In this experiment the bar is used as an indicator (a magnetometer).

FIG. 133.

Now suppose we use the bar as a "source" of a field and hold it fixed so that its long axis is perpendicular to the earth's field (Fig. 133). The induction B produced by this magnetic moment at a point P is (x is large compared to the dimensions of the rod)

$$B = \frac{2\mu_0 m}{x^3} \tag{2}$$

[see Eq. (43b), Chap. V], so that the angle which the resultant of this field and that of the earth makes with the original direction of the earth's field is given by

$$\tan \alpha = \frac{B}{B_0} = \frac{2\mu_0 m}{B_0 x^3} \tag{3}$$

This angle is readily determined with the help of a compass needle. Equations (1) and (3) now allow a simultaneous determination of B_0 and m (choosing μ_0 arbitrarily, *e.g.*, = 1 in e.m.u.). Now we have a calibrated test body which can be used to measure any arbitrary magnetic field. This is the method utilized by Gauss to measure magnetic moments and the earth's magnetic field in an absolute system of units. (H was used instead of B as is done here.)

The utilization of a scalar magnetic potential V_m has already been discussed in Chap. V, and its convenience in describing the field of permanent magnets has been indicated. We shall not pursue the classical development any further but turn now to a discussion of the magnetic properties of matter on the basis of modern atomic ideas.

55. The Electronic Origin of Magnetic Properties.—The fundamental facts concerning the magnetic behavior of material bodies can be most clearly presented by considering a simple experiment. Suppose we consider the magnetic field produced by a closely wound toroidal coil. In empty space this toroid will have a definite inductance, let us say L_0. If now we have this same toroid wound around a material core, it will be found that the inductance of the coil will be different, let us say L, and the ratio L/L_0 is the relative magnetic permeability μ/μ_0 of the medium in which the magnetic field exists when a current flows through the winding. This follows from the fact that the value of H, the magnetic intensity, at a given point within the volume enclosed by this coil is determined in accordance with Ampère's circuital law as independent of the material medium present. This is true in the case of the toroid because the material medium completely fills the region of space in which the magnetic field exists (see Sec. 59). Since the inductance (for a given current) is proportional to the flux of B in this volume, we conclude that, for isotropic homogeneous media, the value of B at every point in the medium has been changed from its value

in empty space in the ratio L/L_0, or since $B = \mu H$, in the ratio μ/μ_0. In contradistinction to the corresponding experiment in electrostatics, in which a dielectric medium is inserted between the plates of a condenser with a consequent *increase* of capacity, the inductance L may be either smaller or larger than the inductance L_0 in the absence of the medium. If $L < L_0$, *i.e.*, $\mu < \mu_0$, we term the medium *diamagnetic* and, if $L > L_0$, *i.e.*, $\mu > \mu_0$, *paramagnetic*. In the diamagnetic case, the effect of the medium is to weaken the field of B relative to its value in empty space (keeping the current in the coil constant), and this is analogous to the dielectric case in electrostatics in which the field intensity $\mathcal{E}$ is weakened by the presence of the dielectric (keeping the charges on the condenser plates constant). There is no electrostatic analogue in this sense to the paramagnetic behavior of material bodies.

In order to understand how diamagnetism and paramagnetism can occur, it is necessary to consider the magnetic properties of atoms (and molecules). As we have already pointed out, an atom consists of a massive positively charged nucleus surrounded by a sufficient number of electrons to make the atom neutral as a whole. These electrons perform some sort of motion around the nucleus, and, when the atom is placed in a magnetic field B, these moving electrons will be acted on by the Lorentz force $-e(v \times B)$, their motions being modified because of this force. The change in the electronic motions caused by the Lorentz force is always such that the altered motion tends to weaken the external field B which gives rise to the change. This is the origin of the *diamagnetic* behavior of atoms and hence of matter. We can see how this behavior comes about with the help of a very simple example. Consider an electron of charge $-e$ moving with a velocity v as shown in Fig. 134. If a magnetic field B into the page is set up, then there will be a deflecting force, as shown acting on the electron, which would result in the circular dotted path shown were the electron free. This is equivalent to a current i flowing as indicated, and this current produces a field which is directed out of the page and hence tends to diminish

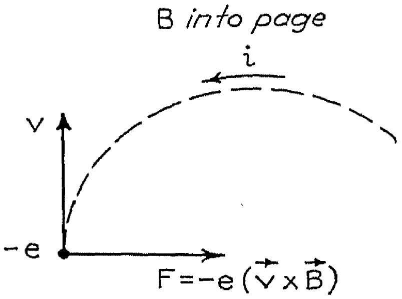

FIG. 134.

the externally applied field. Actually in an atom the electrons are not free but perform orbital motions, so that an electron possesses both angular momentum about the nucleus and a magnetic moment, the latter by virtue of the fact that the orbital motion of a charged particle is equivalent to a tiny current loop. The torque of the Lorentz force on this electron must act at right angles to its angular momentum (since it is a deflecting force) and cannot change the magnitude of the angular momentum, with the result that the angular momentum vector precesses about the direction of the applied field in a manner similar to the precession of a gyroscope. This precession of the electron orbits induced by a magnetic field is called the *Larmor* precession and gives rise to the diamagnetic behavior of the atom.

Besides the magnetic moment possessed by an electron because of its orbital motion, an electron possesses an inherent magnetic moment very much as if it were a spinning sphere of charge and this is called *electron spin*. The resulting magnetic moment of an atom will then be the resultant of the orbital and spin magnetic moments of all the electrons of which it is composed. There are some atoms, *e.g.*, *He* in its normal state, in which there are two electrons and *no* resultant magnetic moment, the magnetic effects of the electrons just neutralizing each other, but many atoms and molecules do possess resultant magnetic moments in their normal states. The diamagnetic effect discussed in the last paragraph will be present whether the atoms have a resultant moment or not; if they have a resultant moment, there is a possibility of another effect due to the tendency of this moment to orient itself in such a direction that its potential energy in the field is a minimum corresponding to stable equilibrium. We have already seen in Chap. V, Sec. 27, that a current loop assumes a stable equilibrium position in the presence of an external field in which the plane of the loop is at right angles to the external field, its magnetic moment in the direction of the external field. In this orientation the field of the current loop aids the external field at the point where the loop is located. This orientation effect gives rise to the *paramagnetic* behavior of material bodies, the magnitude of the natural atomic magnetic moments being large enough to more than compensate for the diamagnetic effect which is always present. The paramagnetic behavior, being an orientation effect, is very much like the

orientation effect of polar molecules in that it is a function of temperature, increasing with decreasing temperature since the thermal agitation of the atoms and molecules tends to hinder the orientation. The diamagnetic effect is essentially temperature independent. One must always keep in mind the fact that inside paramagnetic media the field produced by the natural atomic magnetic moments *aids* the external field, whereas in the dielectric case the field inside the dielectric *always* opposes the external field whether we have induced or oriented dipoles. So much for the qualitative atomic picture.

56. Intensity of Magnetization; Amperian Currents.—In describing quantitatively the magnetic behavior of bodies, it has been customary to introduce the concept of a vector M, the so-called *intensity of magnetization*, which is analogous to the polarization vector P in electrostatics. This vector is defined as the *induced magnetic moment per unit volume* of the magnetized body. In the case of isotropic media, and we shall restrict our attention to this case, the atomic magnetic moments induced (diamagnetic case) or the net result of the orientation of the permanent atomic magnetic moments (paramagnetic case) are opposite to or in the direction of the applied field, so that the intensity of magnetization is a vector in the same direction as the applied field at every point of the medium. Consider a volume element inside a magnetic medium of area ΔA and altitude Δx, as shown in Fig. 135, and let $\vec{m}_0$ be the magnetic moment per atom produced by an externally applied field. The total magnetic moment $\overrightarrow{\Delta m}$ of this volume element is the vector sum of all the atomic $\vec{m}_0$'s in this element, and the intensity of magnetization becomes

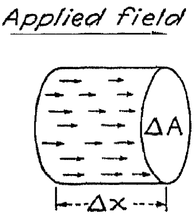

FIG. 135.

$$\vec{M} = \frac{\overrightarrow{\Delta m}}{\Delta v} \tag{4}$$

or more strictly as $\Delta v \longrightarrow 0$

$$\vec{M} = \lim \frac{\overrightarrow{\Delta m}}{\Delta v} \qquad \text{as} \qquad \Delta v \longrightarrow 0 \tag{5}$$

where $\overrightarrow{\Delta m} = \Sigma \vec{m}_0$, the summation extending over all the atoms

in Δv. If the vector field of M is uniform, we say that the substance is uniformly magnetized.

Since each induced magnetic moment m_0 is the equivalent of an elementary current loop, we can equally well attribute the state of magnetization of a body to *circulating currents*, called *Amperian currents* after Ampère, who first suggested them, and these Amperian currents resemble currents in superconductors rather than ordinary currents since their flow involves no dissipation of energy. In describing the magnetic effects of these Amperian currents we make use of the construction of Ampère, as discussed in Sec. 31, Chap. V. For simplicity, consider a cross section of a uniformly magnetized rod as shown in Fig. 136, of thickness Δx, and suppose the direction of the magnetization vector M is into the paper. The induced magnetic moments are equivalent to current loops of area dA, each carrying an equal current Δi_a as shown, the intensity of magnetization being uniform. By Ampère's construction these current loops are equivalent to a surface current Δi_a flowing around the periphery of the bar as shown. To obtain the relation between this induced surface current (the magnetic analogue of the induced surface charges on polarized dielectrics) and the magnetization vector M, we proceed as follows: From the definition of M we have

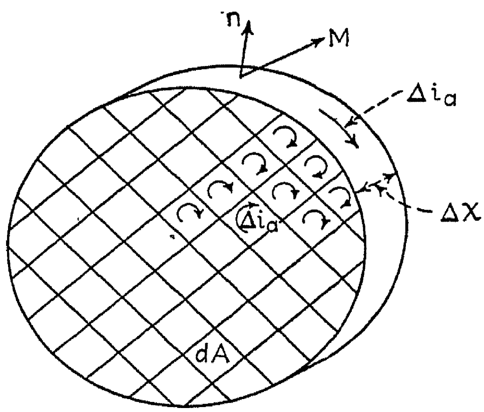

FIG. 136.

$$M = \frac{\Delta m}{\Delta v} = \frac{\Sigma m_0}{A\,\Delta x} \qquad \frac{\qquad dA)}{A\,\Delta x} \qquad \frac{dA}{A\,\Delta x} \qquad \frac{\Delta i_a}{\Delta x} \tag{6}$$

since $\Sigma\, dA$ is the total cross section A of the rod. The expression $\Delta i_a/\Delta x$ is just the *surface density of current*, current per unit length measured along the surface normal to the direction of current flow, which we shall denote by $j_a^{(s)}$, the superscript (s) to remind us that it is a surface current density rather than a volume current density. We thus have the *fundamental relation*

$$j_a^{(s)} = M \tag{7}$$

giving the relation between the magnitudes of the induced surface

current density of Amperian currents and the magnetization vector M. We can include the specification of the directions with the help of a *unit vector* n drawn normal to the surface. From Fig. 136 it is clear that the vector relation

$$j_a^{(s)} = \overrightarrow{M} \times \overrightarrow{n} \tag{8}$$

gives the correct direction of the surface current and it can be shown that Eq. (8) is valid in general. Besides the surface currents required by Eq. (8), one will in general have a volume distribution of Amperian currents in the case of *nonuniform* magnetization, just as one obtained a volume density of polarization charge for nonuniform polarization of a dielectric. One can readily show that the Amperian volume currents are related to the magnetization by the equation

$$\int (j_a)_n \, dS = \oint M_s \, ds \tag{9}$$

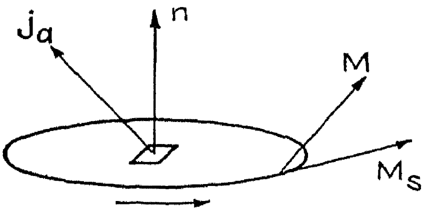

FIG. 137.

i.e., the current crossing any area inside the medium equals the line integral of the magnetization vector around the boundary of the area, the directions as shown in the accompanying figure.

57. Relations of B, H, and M; Magnetic Susceptibility.—In our discussion of magnetic fields in empty space, we found it convenient to introduce a magnetic intensity vector H in addition to the fundamental induction vector B. This was defined essentially as B/μ_0 [see Eqs. (23), (24), and (27), Chap. V], and this simple definition must now be extended for the case of magnetic media, since just for this latter case is this auxiliary vector extremely useful. We shall approach the problem of extending the definition of H with the help of a simple example, *viz.*, a long solenoid of circular cross section, the length large enough compared to the cross-section dimensions so that end effects become negligible, and let us suppose that we insert a cylindrical rod of material coaxially into the interior of the solenoid (Fig. 138). Let the number of turns per unit length of the solenoidal winding be n and the current i. If the cylindrical rod were not present, this would produce a uniform magnetic field inside the solenoid. Thus we see that the rod becomes uniformly magnetized, and we have shown in the last section that the effect of this magnetiza-

tion is equivalent to Amperian currents flowing solenoidally (circumferentially) around the surface of the cylindrical rod. These currents are indicated as dotted in Fig. 138, whereas the "true" current i in the winding is indicated by full lines. We now consider the fundamental induction vector B as produced by *all* the currents, external plus Amperian (just as $\mathcal{E}$ in electrostatics was the field of *all* the charge, true plus polarization) and replace the material rod by the equivalent Amperian currents,

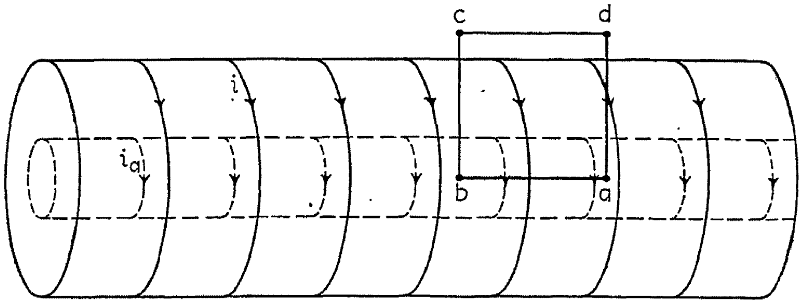

FIG. 138.

thus leaving a problem in empty space which we can readily solve.

We now apply Ampère's circuital law to the path *abcd* in the usual manner, and, remembering that for empty space $H =$ we have (the length *ab* is l) in the usual manner

$$\frac{B}{\mu_0}l = 4\pi(i_{\text{total}}) \tag{10}$$

where i_{total} is the total current (including the Amperian currents) passing through the area *abcd*. We have for the total current

$$i_{\text{total}} = nil + j_a^{(s)}l = (ni + M)l \tag{11}$$

utilizing Eq. (7). Thus Eq. (10) yields

$$\frac{B}{\mu_0} = 4\pi ni + 4\pi M$$

or

$$\frac{B}{\mu_0} - 4\pi M = 4\pi ni \tag{12}$$

Thus we see that the quantity $4\pi ni$, which is a measure of H inside such a solenoid, is no longer equal to B/μ_0 when matter is present, and we *define* H by the more general relation

$$H = \frac{B}{\mu_0} - 4\pi M$$

This reduces to the previous case if M is zero, as in the case of empty space. We see in this example that H is the field of the external currents, and B the field of the total current, both external and hidden (Amperian).

The general problem of determining the magnetic induction B thus involves a knowledge of both H and M. The magnetization vector M, however, is in general a function of H (or B), and consequently we must distinguish among various cases which are actually found in material bodies. Let us restrict ourselves for the present to those materials in which B, H, and M are all proportional to each other, the magnetization being proportional to the field. Calling the ratio of B to H the magnetic permeability of the medium, $\mu = B/H$, we have from Eq. (13)

$$\frac{\mu}{\mu_0} = 1 + 4\pi\frac{M}{H} \tag{14}$$

and the ratio M/H is called the *magnetic susceptibility* of the medium and is denoted by χ_m. For most materials (the notable exceptions being iron, nickel, cobalt, and other so-called ferromagnetic substances) the intensity of magnetization is proportional to H, and χ_m is a constant. Materials for which χ_m is negative are called *diamagnetic* ($\mu < \mu_0$), and those for which χ_m is positive are termed *paramagnetic* ($\mu > \mu_0$). For paramagnetic media the Amperian currents aid the external currents, and in diamagnetic media they oppose them. In terms of the susceptibility, Eq. (14) can then be written

$$\frac{\mu}{\mu_0} = 1 + 4\pi\chi_m \tag{15}$$

The susceptibilities of ordinary paramagnetic and diamagnetic substances are very small compared to unity, being of the order of 10^{-6} for diamagnetic bodies (bismuth is a notable exception with about ten times this susceptibility) and somewhat larger, of the order of 10^{-5}, let us say, for paramagnetic bodies. Thus it is possible to treat most substances as being nonmagnetic for the purposes of many practical problems.

58. Ferromagnetism.—From a practical standpoint, by far the most important magnetic media are the so-called *ferromagnetic*

materials which are characterized by abnormally large values of the magnetization M and by the fact that the magnetization is *not* proportional to H; indeed in some substances it is not a single-valued function of H. The elements iron, nickel, cobalt, and a number of alloys display this abnormally large paramagnetic behavior. Equation (13) still holds, since it is a definition, but the permeability defined as B/H is at best a function of H. Not only is the magnetization intensity much larger for these substances than for ordinary paramagnetic materials (sometimes a million times larger), but also it is possible to attain a limiting saturation value of M at relatively low field strengths. The saturation value of M is relatively independent of the mechanical state and small amounts of impurities, but the B-H or M-H relation (the so-called magnetization curve) is very strongly dependent on these factors. It is convenient to classify ferromagnetic materials into two groups:

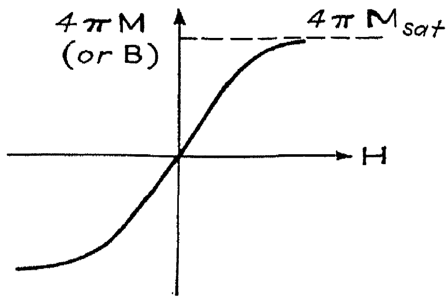

Fig. 139.

a. Magnetically Soft Substances.—These are substances for which M is at least approximately a single-valued function of H. This function is shown in Fig. 139 and has the following general characteristics; an initial sharp rise in M and a later flattening out and saturation. Actually there are no strictly reversible ferromagnetics, but one can only speak of "softer" or "harder" magnetic substances depending on the size of the hysteresis loop (see below).

b. Magnetically Hard Substances.—In these the magnetization intensity not only is not a single-valued function of H but depends on the previous history of the sample under consideration. If one subjects an initially unmagnetized sample of magnetically hard steel to an increasing magnetic field H, the initial magnetization curve (shown dotted in Fig. 140) is not unlike that of Fig. 139. If now the applied field is reduced and reversed, the magnetization follows the solid curve PAQ. The value of M for $H = 0$, the ordinate OA in Fig. 140 (a measure of the so-called "remanence" of the substance), and the reversing field OB necessary to reduce M to zero (the so-called "coercive force)" can be used as measures of the magnetic hardness of the material. If one now carries the substance back to the point P

by increasing H, the lower curve is followed, and this cyclic operation is exactly what occurs in a.c. transformers. The loop $PAQCP$ is called a "hysteresis" loop, and there is an energy loss whenever such a loop is traversed. Consider a toroidal coil closely wound around a steel core of cross section A and mean length l. When a current is set up in the magnetizing coil, energy is supplied to the system at the rate Ei, where E is the voltage across the coil and i the current through it. Neglecting the resistance of the coil (taking this into account would simply add the ordinary i^2R heating), this voltage is given by Faraday's law as

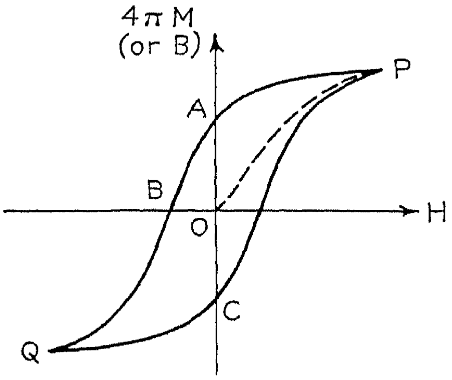

FIG. 140.

$$E = NA\frac{dB}{dt}$$

where N is the number of turns of the coil. The value of H is related to the current i by

$$H = \frac{4\pi Ni}{l}$$

so that

$$i = \frac{Hl}{4\pi N}$$

and the rate of doing work is

$$Ei = \frac{Al}{4\pi}H\frac{dB}{dt} = \frac{v}{4\pi}H\frac{dB}{dt}$$

where v is the volume of the steel specimen. The total work done in carrying the substance around the hysteresis loop is hence

$$W = \int Ei\,dt = \frac{v}{4\pi}\int H\frac{dB}{dt}dt = \frac{v}{4\pi}\oint H\,dB \qquad (16)$$

so that the hysteresis loss per cycle per unit volume is $1/4\pi$ times the area of the loop on a B-H diagram.

In the interior of a permanent magnet the direction of the magnetic intensity H is generally opposite to that of the magnetization and induction. Such a magnetic state corresponds to a position such as is found on the portion of the curve of Fig. 140 lying between A and B. Furthermore, the internal field of such a magnet depends, for a given magnetization, on the geometrical shape of the magnet.

59. Boundary Conditions on B and H.—Up to this point in our discussion of the magnetic behavior of material bodies, we

have confined ourselves either to the case of a single material medium completely occupying the region of space in which the magnetic field existed, or to the case where the boundary surface between two media (the case of Fig. 138) was everywhere parallel to the direction of the magnetic field which existed before the insertion of the material body. In both these cases the direction of the lines of B or H is unaltered by the presence of the material body, and we were not concerned with the possibility of the refraction of these lines at the interface. We must now investigate the relations which hold at a boundary which is not parallel to the lines of force, let us say a boundary surface between two media of permeabilities μ_1 and μ_2. Since the field of B is solenoidal (the lines of B closing on themselves) Gauss's theorem, applied to the flux of this vector, states that the net flux of B emerging from any closed volume must be zero. We apply this theorem to the shallow pillbox shown in Fig. 141, and we may neglect the flux of B emerging from the curved sides, since it vanishes as the altitude of the pillbox approaches zero (the boundary surface always lying between the flat faces). The flux of B emerging from the top face is $B_{2n}\,dA$, where B_{2n} is the normal component of B_2 at the point where this surface element dA is located. Similarly the emergent flux across the bottom face is $-B_{1n}\,dA$. Hence, we have

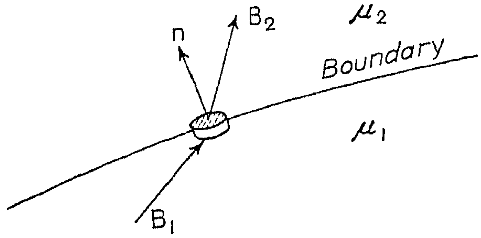

FIG. 141.

$$B_{2n}\,dA - B_{1n}\,dA = 0 \tag{17}$$

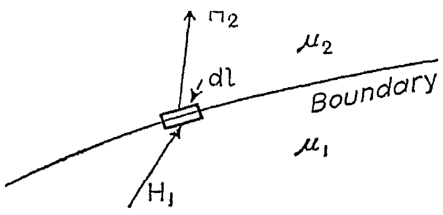

FIG. 142.

The normal component of B is continuous at the boundary surface between two media. Note that this condition is automatically satisfied if the boundary is parallel to the lines of B.

For the tangential components of the field we make application of the Ampère circuital law and choose a closed path as shown in Fig. 142, in which the sides perpendicular to the boundary can be made vanishingly small compared to the other sides, each of length dl. As this occurs, the displacement current through the

area enclosed by this path becomes vanishingly small and does not contribute to the magnetomotive force around the closed path. This magnetomotive force is given by

$$\oint H_s\,ds = H_{1t}\,dl - H_{2t}\,dl$$

and this must equal 4π times the true current flowing across the area enclosed by our elementary path. If the surface density of this current on the interface is $j^{(s)}$, we have

$$H_{1t}\,dl - H_{2t}\,dl = 4\pi j^{(s)}\,dl$$

or

$$H_{1t} - H_{2t} = 4\pi j^{(s)} \tag{18}$$

as the relation which must be satisfied by the tangential components of H. If no true surface currents are present (there may be Amperian currents, however), $j^{(s)} = 0$, and Eq. (18) becomes

$$H_{1t} = H_{2t} \tag{19}$$

establishing the continuity of the tangential components of H for this case.

Using the relations

$$B_1 = \mu_1 H_1; \qquad B_2 = \mu_2 H_2$$

one finds readily the law of refraction of the lines of B or H at an interface which carries no true surface current. It is

$$\frac{\tan\alpha_1}{\tan\alpha_2} = \frac{\mu_1}{\mu_2} \tag{20}$$

where α_1 and α_2 are the angles which H or B makes with the normal to the surface in media 1 and 2, respectively. This law is entirely similar to the corresponding law for the electric field.

We are now in a position to state precisely the conditions under which Ampère's rule [Eq. (24), Sec. 28, Chap. V] may be employed to calculate the field of H and thereby bring out clearly the generality of the Ampère circuital law, which is always valid, over the Ampère rule. From our discussion of the boundary conditions it becomes clear that Ampère's rule will give correctly the field of H in two cases, (1) when the region of space in which the magnetic field exists is completely occupied by a single homogeneous material medium, or when the boundaries

between media are so far from the region of space under consideration that they produce no sensible effect in this region, and (2) when the boundary surface between bodies is everywhere parallel to the magnetic lines, so that the field pattern is independent of the presence of the boundary. This follows from the facts that Eq. (17) is automatically satisfied and Eq. (19) insures an unchanged value of H as one crosses the boundary. Only in these cases may we say that H is completely and uniquely determined by the true currents and their relative positions, whereas the magnetomotive force around *any* closed path is always given uniquely by the true current (including displacement current) which traverses any surface bounded by this path. We can illustrate these statements with the help of a simple example. Consider the magnetic field produced by a very long solenoid in air carrying a steady current. Near the center of this solenoid the field is confined to the region of space enclosed by the winding and is uniform in this region (see Sec. 29, Chap. V). If a cylindrical rod of magnetic material is inserted coaxially into this solenoid, either filling or only partially filling it, the field of H is exactly as it was before (still confining our attention to the central portion of the solenoid), and the pattern of magnetic lines is unchanged. The only change which occurs is the change in the value of B inside the magnetic material. Now, however, let us imagine that we insert a short cylinder of this magnetic material coaxially into the central part of the solenoid. The whole field pattern changes violently and is neither uniform in the central part of the solenoid nor is it confined to the space enclosed by the solenoidal winding. We can obtain a good qualitative picture of the field in this case with the help of the concept of Amperian currents, replacing the material cylinder by a solenoid of equal length. The number of ampere turns of this short solenoid may be enormous if the material is ferromagnetic. The field pattern will now be that obtained by superposing the field produced by this finite "Amperian" solenoid (compare Fig. 65, page 107) and that produced by the original very long solenoid in air. It becomes evident that the presence of the flat surfaces of the rod (or the ends of the "Amperian" solenoid) play an important part in modifying the original field pattern.

Let us investigate more closely the effect of these boundaries in a very simple case. Suppose we consider the case where the lines

of B or H are normal to the interface between two media, as shown in Fig. 143. In this figure we have indicated the lines of H in the two media, assuming that $\mu_2 > \mu_1$. Since the field is normal to the boundary, Eq. (19) is satisfied and Eq. (17) demands that

$$\mu_1 H_1 = \mu_2 H_2$$

or

$$H_1 = \frac{\mu_2}{\mu_1} H_2$$

so that the number of lines of H emerging per unit area of the interface into medium 1 is larger than the number incident per unit area on the interface in medium 2. This discontinuity is not present when we consider the lines of B, since $B_1 = B_2$. Whenever new lines of force start (or stop) at definite points of space, it is natural to think of these new lines as originating in "sources" of the field at these points. Thus we may say that it is possible to have sources of the field of H but not of B. These sources of H have been called *magnetic poles*, and, before it was recognized that B is the fundamental magnetic vector and H merely an auxiliary aid to calculation, real physical significance was given to these poles. It seems better, in the light of our present knowledge of the subject, to look upon them merely as possible modes of description and indeed they are very convenient concepts when applied to many engineering problems, especially those in which one is interested in the magnetic field in an air gap which has been introduced in an otherwise closed "magnetic circuit" (see below) composed of a ferromagnetic substance of high permeability.

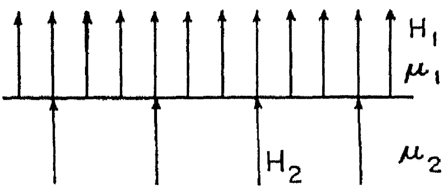

FIG. 143.

One can set up a measure of the "strength" of a magnetic pole in terms of the number of lines of H produced by it. For example, suppose we consider an interface area A in Fig. 143, and let us suppose that the medium 1 is empty space, so that $\mu_1 = \mu_0$ and that medium 2 is a ferromagnetic body. The number of lines of H starting at the pole on this surface A is clearly

$$(H_1 - H_2)A$$

and, since we have the relations

$$H_1 = \frac{B}{\mu_0}$$

and

$$H_2 = \frac{B}{\mu_0} - 4\pi M$$

where M is the magnetization intensity at the surface, the expression for the strength of the magnetic pole on this surface is

$$\text{Pole strength} = (H_1 - H_2)A = 4\pi MA \tag{21}$$

or alternatively,

$$\frac{\text{Pole strength}}{\text{Unit area}} = 4\pi M$$

Clearly one might adopt units of pole strength such that it equals $1/4\pi$ times the net flux of H emerging from the surface, in which case the pole strength per unit area would be simply equal to the intensity of magnetization at the surface. This has been the common definition. More precisely, the M in Eqs. (21) and (22) is the normal component of M at the surface, as is evident from our derivation. One fact becomes very clear from these considerations, namely, that the constancy of magnetic pole strength implies constancy of the magnetization intensity, and only for substances which are so hard magnetically that the remanent magnetization is practically independent of the external field is it possible to assign even an approximate meaning to the term pole strength as a property of the system independent of its external surroundings. In electrostatics, however, the electric charge on an insulated body is strictly constant and thus can be used uniquely to detect and measure electric fields. With magnetic poles we are never quite sure of our basic assumption, for even in weak fields, immersion of a permanent magnet in a medium of high permeability will certainly modify the pole strength by virtue of the "induced" magnetization and the consequent induced poles. For reasons such as these we have preferred to introduce the magnetic field concept on the basis of electric current rather than from the standpoint of permanent magnets, as is commonly done.

60. Magnetic Circuits; Reluctance.—The determination of the field of magnetic induction in the presence of arbitrary magnetic

bodies is, in general, an extremely difficult task and far beyond the scope of this book. The general problem for steady fields involves the simultaneous solution of the equations

$$\oint H_s\, ds = 4\pi i$$

$$\int_{\substack{\text{closed}\\\text{surface}}} B_n\, ds = 0$$

$$B = \mu H$$

subject to the boundary conditions expressed by Eqs. (17) and (19). There are, however, certain problems of practical importance, involving so-called *magnetic circuits*, for which it is possible to readily obtain approximate solutions. The name magnetic circuit has its origin in certain analogies between this sort of problem and that of the flow of steady currents in linear conductors. The fundamental reason for the analogy lies in the fact that both the current-density field and the magnetic induction field are solenoidal, there being no sources or sinks, and the lines are closed on themselves in both cases. If one compares the equations

$$j = \sigma \mathcal{E} \qquad \text{and} \qquad B = \mu H$$

(the first is Ohm's law), one sees that in a sense μ can be looked upon as the analogue of the conductivity σ. Now, in simple electrical problems, the current can be easily confined to conducting bodies. In particular for linear conductors, such as wires, the lines of current flow are parallel to the boundary surfaces of the wires and are uniformly distributed across the cross section if the latter is uniform. For this case, Ohm's law applied to a simple series circuit takes the more convenient form

$$E = \oint \mathcal{E}_s\, ds = iR,$$

where the resistance R is related to the conductivity by

$$R = \frac{l}{\sigma A}$$

l being the length of the conductor and A its cross section.

In the corresponding magnetic circuit one has a closed path of highly permeable material such as iron, as shown in Fig. 144,

with an exciting winding of N turns carrying a steady current i. We make the following assumptions:

1. The lines of B are confined to the circuit and are parallel to the boundaries of the magnetic medium. The higher the ratio μ/μ_0 and the smaller the cross section A relative to the length l of the circuit, the more nearly is this fulfilled.

2. The magnetic permeability μ is constant. This is not nearly so true as the corresponding statement that σ is a constant (at constant temperature) for a conductor. We shall use μ to denote a mean value.

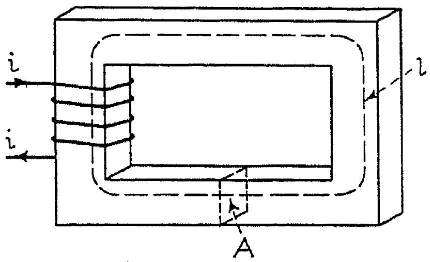

FIG. 144.

3. The values of H and B over any cross section of the circuit may be replaced by mean values over such a cross section. This involves a choice of mean length l of the circuit which, at least in practical cases, has not a very small ratio of cross-section dimension to length.

With these assumptions we can set up equations analogous to those describing the corresponding electrical circuit. The magnetomotive force around the path l is, using Ampère's circuital law,

$$\oint H_s \, ds = Hl = 4\pi Ni$$

where now H is the mean value of H in the medium and l the mean length. From this we obtain the magnetic flux Φ as

$$\Phi = BA = \mu HA = \frac{4\pi Ni\mu A}{l}$$

or rewritten

$$\Phi = \frac{4\pi Ni}{l/\mu A} = \frac{\text{m.m.f.}}{\mathcal{R}} \tag{23}$$

where $4\pi Ni$ is the magnetomotive force around the circuit (m.m.f.) and the script $\mathcal{R}$ is called the *reluctance* of the magnetic circuit. The reluctance is the analogue of electrical resistance as we see by writing Ohm's law for a corresponding simple series circuit.

$$i = \frac{\text{e.m.f.}}{R}$$

From the similarity of the equations one can readily see that, subject to the same assumptions, the analysis of two or more reluctances in series or in parallel can be handled by the same method. We shall illustrate this for the case of a simple electromagnet, as shown in Fig. 145, in which there is an air gap of length d. Denoting by H the value of H inside the iron and by H_0 its value in the air gap, we have for the magnetomotive force around the path $(l + d)$,

$$\text{m.m.f.} = \oint H_s\,ds = Hl + H_0 d = 4\pi Ni$$

If we assume that the effective air-gap area is equal to A, that of the iron (this neglects fringing and is a good approximation

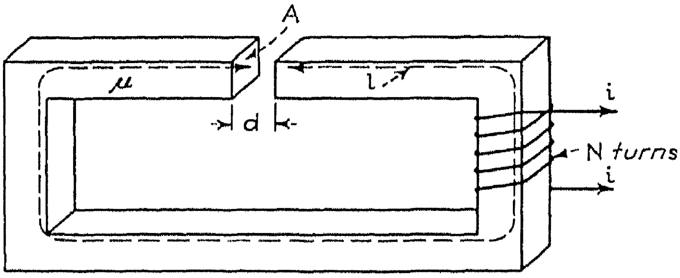

FIG. 145.

only if $d^2 \ll A$), we can set the magnetic flux Φ equal to BA. From the above equation we have

$$B\left(\frac{l}{\mu} + \frac{d}{\mu_0}\right) = 4\pi Ni$$

using $B = \mu H$ and the fact that B is continuous at the boundaries, so that

$$\Phi = BA = \frac{4\pi Ni}{(l/\mu A) + (d/\mu_0 A)} = \frac{\text{m.m.f.}}{\mathcal{R} + \mathcal{R}_0}$$

where $\mathcal{R}$ is the reluctance of the iron path and $\mathcal{R}_0$ is the reluctance of the air gap. Thus we see that reluctances in series add just as do resistances. The magnetomotive force across the gap is given by

$$\mathcal{R}_0\Phi = \frac{4\pi Ni}{1 + (l\mu_0/d\mu)}$$

and if the ratio l/d is 100, let us say, and the permeability of the iron is $2{,}000\mu_0$, then this magnetomotive force is $\frac{20}{21}$ of the total

ı.f. around the circuit, *i.e.*, about 95 per cent of the 'drop" cross the air gap.

Problems

1. A very long solenoid having 20 turns per centimeter is wound on an core 3 cm. in radius and a current of 10 amp. flows through the winding. permeability of the iron (assumed constant) is $2{,}000\mu_0$. Neglecting effects, calculate:

. The self-inductance per centimeter length of the solenoid.

. The intensity of magnetization inside the iron core.

. The induced Amperian current per centimeter flowing solenoidally ınd the core surface.

. The number of ampere turns per centimeter for a solenoid in air of ilar dimensions needed to produce the same inductance.

. The magnetic field energy stored per centimeter length of the solenoid.

2. A toroidal coil has a mean radius of 10 cm. and a cross section of m^2. It has 1,500 turns wound on a core of permeability $800\mu_0$. If the stance of the winding is 2 ohms, compute the time constant of the coil. sume the field uniform over a cross section of the core.)

3. Explain how the cavity definitions of D and $\mathcal{E}$ in electrostatics may carried over to the magnetic case, showing that the field inside a needle- ped cavity with its long axis parallel to the direction of magnetization is en by H, whereas that inside a pillbox cavity with its faces normal to magnetization direction is B.

4. A long straight copper wire 1 cm. in diameter is surrounded coaxially a long hollow iron cylinder of permeability $1{,}000\mu_0$, inner radius 2 cm., d outer radius 3 cm. The wire carries a steady current of 20 amp.

a. Compute the total magnetic flux inside a section of the iron cylinder neter long.

b. The induced Amperian currents flowing on the surfaces of the iron linder are parallel or antiparallel to that in the copper wire and are uni- rmly distributed over these surfaces. Compute the magnitudes of these rrents on both inner and outer surface of the hollow cylinder and their rections relative to the current in the copper wire.

c. Compute the intensity of magnetization at a point 2.5 cm. from the is of the copper wire.

d. Prove that the Amperian current density inside the iron is zero.

e. Show that the magnetic field outside the iron cylinder is the same as if e iron were absent.

5. A very long solenoid of radius R is wound with n turns per unit length ıd a long cylindrical rod of radius $r < R$ is placed coaxially inside the lenoid.

a. If the permeability of the rod material is μ, derive an expression for ıe self-inductance per unit length of the solenoid.

b A coil of N turns is wound around the cylindrical rod, and an alternating rrent $I \sin \omega t$ flows in the outside winding. Derive an expression for the m.f. induced in the secondary coil of N turns, assuming it to be on open rcuit.

6. A 1,200-turn toroid is wound on an iron ring of mean diameter 18 cm. and cross section 6 cm.^{2}, and a current of 2 amp. flows in the winding. The permeability of the iron is $3,000\mu_0$.

a. Compute the flux of B in the ring.

b. If an air gap of length 0.5 mm. is cut in the ring, compute the flux in this air gap, assuming its effective area to be that of the ring.

c. Compute the inductance of the coil with and without the air gap.

d. Calculate the total field energy when there is no air gap.

e. Calculate the total field energy and the field energy in the iron and in the air gap.

7. An iron rod of square cross section (2 by 2 cm.), of relative permeability 1,600, is bent into the form of a ring of inner radius 5 cm., and the ends are welded together. Wire is wound toroidally around the ring to form a coil of 500 turns and a current of 1 amp. flows through the winding.

a. Compute the total flux of B in the ring, taking into account the variation of B inside the iron. What is the inductance of the coil?

b. What is the magnitude of the Amperian current flowing around the surface of the ring?

c. What is the maximum percentage variation of the magnetization intensity inside the iron? Where is the magnetization largest and where is it smallest?

d. Prove that the density of Amperian currents is zero everywhere inside the iron.

8. A 3-mm. air gap is cut in the ring of the toroid of Prob. 7.

a. Compute the flux of B inside the ring and the self-inductance of the coil.

b. A square slab of iron (2 by 2 cm.), of thickness 2 mm. and relative permeability 2,400, is inserted into the air gap so that the edges of one of its faces coincide with those of the toroidal core. How much work must be done by the sources which maintain the current in the coil constant during the insertion process?

c. What is the increase of magnetic field energy during the process described in part *b*, and how much mechanical work is done by the force which pulls the slab into the gap?

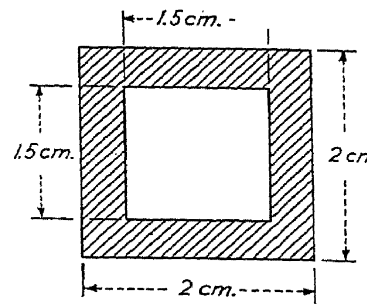

FIG. 146.

9. Suppose the iron rod used in constructing the toroid of Prob. 7 is hollow and has the cross section shown in Fig. 146. Compute the self-inductance of the toroid in henrys, taking into account the variation of B with position.

10. Prove that the magnetic field energy stored in a magnetic circuit can be written as

$$U_m = \frac{1}{8\pi}\Phi^2\mathcal{R} = \frac{1}{8\pi}(\text{m.m.f.})\Phi = \frac{1}{8\pi\mathcal{R}}(\text{m.m.f.})^2$$

where Φ is the total flux of B in the circuit, $\mathcal{R}$ the reluctance, and (m.m.f.) is the magnetomotive force around the circuit.

11. Consider the magnetic circuit of uniform cross section A shown in Fig. 147. The air-gap lengths are very small compared to the cross-section dimension.

a. Using the results of Prob. 10, show that, if the air gaps are closed, the field energy increases by

$$\frac{(\text{m.m.f.})^2}{8\pi}\left(\frac{1}{\mathcal{R}_2} - \frac{1}{\mathcal{R}_1}\right)$$

where $\mathcal{R}_2$ is the reluctance of the circuit without air gaps and $\mathcal{R}_1$ its value with the air gaps. The current in the winding is kept constant.

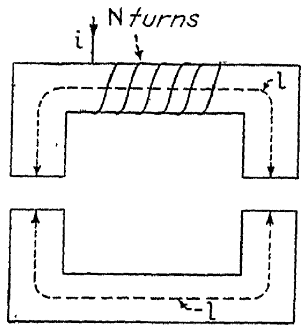

FIG. 147.

b. Using Faraday's induction law, show that the sources of e.m.f. maintaining the current constant do an amount of work on the system equal to twice the above expression while the air gap is being closed. What happens to the difference of these energies?

12. Starting with the magnetic circuit shown in Fig. 147, suppose the top half is held fixed and the bottom half is allowed to move up an infinitesimal distance dx. Derive an expression for the increase of field energy during this displacement, keeping the current in the magnetizing coil constant, and, using the results of Prob. 11, show that the force with which the two sections attract each other is given by

$$F = 2\frac{B^2A}{8\pi\mu_0}$$

13. Two long iron plungers of permeability 2,000 are inserted into a very long solenoid, the plungers each of 4 cm.^{2} cross section and fitting tightly in the solenoid. If the magnetic induction in the iron is 5,000 gauss, compute the force in pounds with which one must pull to separate the plungers.

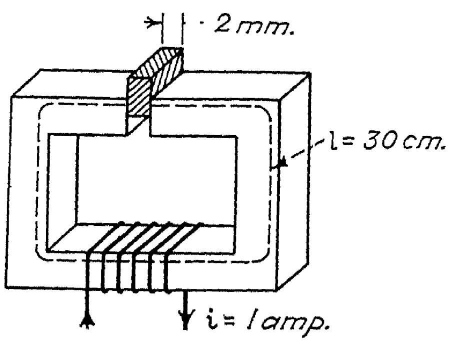

FIG. 148.

14. A long solenoid of 10 turns per centimeter contains an iron rod, 2 cm. in diameter cut in two, and carries a current of 3 amp. in its winding (the primary). Compute the force necessary to separate the two halves if the following experimental data are known: On reversing the current in the primary a charge of 60 microcoulombs flows through a secondary circuit consisting of a coil of 10 turns and 100 ohms resistance, wound on the same core.

15. A magnetic circuit has a core of mean length 30 cm., cross section 6 by 6 cm., and an air gap 2 mm. long, as shown in Fig. 148. The magnetizing coil has 100 turns and carries a steady current of 1 amp. A slab of iron 6 by 6 by 0.2 cm. is inserted into the air gap as shown. Assuming the relative permeability of the core mate-

rial and slab to be constant and equal to $1{,}200\mu_0$, compute the work done in pulling the slab into the air gap.

16. Consider a region of space in which there exists a magnetic field which is not quite uniform. A rigid magnetic body of permeability μ, volume v, is brought into the field at a point where the magnetic induction was B. Suppose the volume v of the body is small enough so that B does not vary appreciably throughout this volume and that the susceptibility of the body is extremely small compared to unity (this is always true of ordinary paramagnetic and diamagnetic media), so that we may consider the induction B unaltered by the presence of the body. Show that the decrease of magnetic field energy caused by this insertion is given very nearly by

$$\frac{v}{8\pi}(\mu - \mu_0)\frac{B^2}{\mu_0^2} = \frac{v}{8\pi}(\mu - \mu_0)H^2$$

where H is the magnetic intensity at the point where the body is located prior to its introduction.

Now consider a small displacement of the body to a point where the magnetic field has a slightly different value. Compute the decrease in field energy due to this displacement, and, equating this to the work done by mechanical forces on the body, show that the mechanical force F acting on the magnetic body is given by

$$F = \frac{v}{2}\chi_m\mu_0 \text{ grad } H^2$$

where χ_m is the magnetic susceptibility of the body.

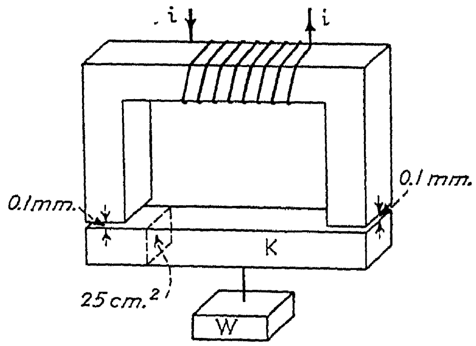

FIG. 149.

17. An electromagnet as shown in Fig. 149 is designed to support a weight of 100 lb. (including the weight of the keeper K). The cross section of all parts of the magnetic circuit is 25 cm.², the length of the magnetic path in the iron (including the keeper) is 50 cm., and the air gaps are each 0.1 mm. long. The permeability of the iron is $1{,}800\mu_0$. If the wire of the magnetizing coil can carry 1 amp., compute the least number of turns needed in the coil to support this weight.

18. Suppose that the iron in Prob. 17 has not a constant permeability but that μ varies with B according to the following table:

B, gauss	3,500	4,000	4,400	4,900	5,300	5,800
μ/μ_0	1,650	1,600	1,560	1,500	1,450	1,380

Compute the least number of turns needed in the coil, assuming a maximum current of 1 amp.

19. A magnetic core of constant length and uniform cross section has an adjustable air gap. On this core there is wound a coil. When the air gap is reduced to zero, the self-inductance of the coil is 3 henrys. When the length of the air gap is 0.05 in., the inductance is 1 henry. To what length must the air gap be increased to reduce the inductance to 0.2 henry? Assume a constant permeability of the magnetic core.

20. Two coils are wound side by side on the same magnetic circuit. The first has 500 turns, a resistance of 2.5 ohms, a self-inductance of 0.5 henry; and the second has 1,000 turns and a resistance of 50 ohms. If an e.m.f. of 4.3 volts (direct current) is impressed on the first coil, what voltage must be impressed on the other coil so that the total flux in the core will be reduced to zero? What will be the flux in the core when a voltage of 15 volts is impressed on the second coil alone?

21. Starting from the fact that

$$\frac{1}{\mu_0}\oint B_s\, ds = 4\pi(i_{\text{total}})$$

where i_{total} is the sum of the true and Amperian currents flowing across any surface of which the closed path of integration is a boundary, prove, with the help of the general definition of H, that

$$\oint M_s\, ds = i_a = \int (j_a)_n\, dS$$

where i_a is the Amperian current flowing across the above mentioned surface and j_a is the volume density of Amperian currents. M is the intensity of magnetization.

22. Consider the volume enclosed by a hemispherical surface in a region where a magnetic field changes with time. Compare the expressions for the e.m.f. induced around the equator of the hemisphere by considering (a) the magnetic flux crossing the plane surface of the hemisphere and (b) the magnetic flux crossing the curved surface. Equating these expressions (they both express the same e.m.f.), prove that

$$\int_{\substack{\text{closed}\\ \text{surface}}} B_n\, dS = 0$$

i.e., that the field of B, not of H, is solenoidal.

CHAPTER XIII

ELECTROMAGNETIC WAVES IN MATERIAL BODIES

In Chaps. VIII and IX we have seen how the introduction of the concept of displacement current into the fundamental laws of electromagnetic theory led to the prediction of the possibility of electromagnetic waves, traveling in empty space with a velocity

$$c = \frac{1}{\sqrt{\epsilon_0\mu_0}} = 3 \times 10^{10} \text{ cm./sec.} \qquad \text{(the velocity of light)}$$

We have shown that, at least for plane waves, the waves are transverse, both $\mathcal{E}$ and H having no components in the direction of propagation, and that $\mathcal{E}$ and H are perpendicular to each other. The energy transported by these waves can be expressed by means of the Poynting vector $S = (c/4\pi)(\mathcal{E} \times H)$, this vector giving the direction of propagation as well as the intensity. In this chapter we shall study the behavior of such waves when they are propagated in material media, especially in dielectrics, and in particular we shall examine their behavior as they impinge on a boundary surface which separates two dielectric media. According to the electromagnetic theory of light, we should expect that the laws so found will be valid in describing optical phenomena, and we shall concern ourselves largely with applications to the field of optics. Throughout this chapter, and in all our discussion of optical phenomena, we shall use *Gaussian units* exclusively. Reformulation of the laws in m.k.s. units (or any other) is left as an optional task for the student.

61. Plane Waves in Dielectrics.—We start with a discussion of electromagnetic waves in uncharged nonconducting stationary bodies for which the fundamental laws governing the electric and magnetic field vectors may be written in exactly the same form as we wrote them for empty space, since the density of true charge ρ and the real current density j vanish in both cases. In Gaussian units these laws may be written in the form

$$\int_{\substack{\text{closed}\\ \text{surface}}} D_n\,dS = 0; \qquad \int_{\substack{\text{closed}\\ \text{surface}}} B_n\,dS = 0 \tag{1}$$

$$\oint \mathcal{E}_s\,ds = -\frac{1}{c}\int \frac{\partial B_n}{\partial t}dS; \qquad \oint H_s\,ds = \frac{1}{c}\int \frac{\partial D_n}{\partial t}dS \tag{2}$$

In addition to these we have the relations $D = \epsilon\mathcal{E}$ and $B = \mu H$. For all but ferromagnetic bodies—and these are conductors—the magnetic susceptibility is so extremely small compared to unity that we may take $\mu = 1$ (e.m.u.) without appreciable error. We shall, however, write most of our equations in a general form, retaining an arbitrary value of μ.

Let us first review briefly the arguments leading to the equation for linearly polarized waves traveling along the x-axis. We have seen that Eqs. (1) demand that the x-components of all the vectors be zero (or at the most be constant and hence not of interest for the study of waves) and that both $\mathcal{E}$ and H (also B and D) must lie in planes normal to the x-axis. Equations (2) were then applied to the elementary circuits shown as I and II in Fig. 150. The first of Eqs. (2) applied to circuit I led to the relation

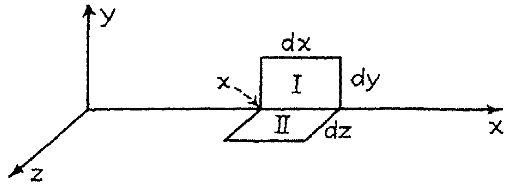

FIG. 150.

$$\frac{\partial \mathcal{E}_y}{\partial x} = -\frac{1}{c}\frac{\partial B_z}{\partial t} \tag{3}$$

and the second of Eqs. (2) applied to circuit II gave

$$\frac{\partial H_z}{\partial x} = -\frac{1}{c}\frac{\partial D_y}{\partial t} \tag{4}$$

These equations are to be supplemented by two more in the general case of plane waves traveling along the x-axis with arbitrary polarization, and these are

$$\frac{\partial \mathcal{E}_z}{\partial x} = +\frac{1}{c}\frac{\partial B_y}{\partial t}; \qquad \frac{\partial H_y}{\partial x} = +\frac{1}{c}\frac{\partial D_z}{\partial t} \tag{5}$$

although we shall not need to make explicit use of these last relations. The only difference in the argument from that of Chap. IX is now to replace B by μH (instead of $\mu_0 H$) and D by

$\epsilon\mathcal{E}$ (instead of $\epsilon_0\mathcal{E}$). There then follows the wave equation for $\mathcal{E}_y$,

$$\frac{\partial^2\mathcal{E}_y}{\partial x^2} = \frac{\epsilon\mu}{c^2}\frac{\partial^2\mathcal{E}_y}{\partial t^2} \tag{6}$$

with an identical equation for H_z. If we consider linearly polarized waves, then only $\mathcal{E}_y$ and H_z are different from zero, all other components of these vectors being zero. Equation (6) may be written in the form

$$\frac{\partial^2\mathcal{E}_y}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2\mathcal{E}_y}{\partial t^2} \tag{6a}$$

where $v = c/\sqrt{\epsilon\mu}$ is the velocity of the wave which is now different from its velocity c in empty space. In the case of nonmagnetic bodies we may set $\mu = 1$ and find the relation

$$v = \frac{c}{\sqrt{\kappa}} = \frac{c}{n} \tag{7}$$

where κ is the dielectric constant of the medium. In optics it is usual to denote the ratio of the velocity of light in empty space to its value in a material medium as the *index of refraction* n of the substance. Hence we predict the relation

$$n = \sqrt{\kappa} \tag{8}$$

between the index of refraction and dielectric constant of a dielectric. Experimentally it turns out that this equality is *not* true in general, and there are violent exceptions. For example, water has an index of refraction of about 1.3, whereas $\sqrt{\kappa} = 9$. The reason for the discrepancy lies, not in the inadequacy of the fundamental laws of electromagnetic theory, but rather in the tacit assumption that the dielectric constant of a dielectric is strictly constant, independent of frequency. This assumption is not justified when dealing with waves of optical frequencies, and we shall investigate the theory of the variation of κ with frequency in a later chapter on the dispersion of light, *i.e.*, the variation of index of refraction with frequency (or wave length). Equation (6*a*) is satisfied by traveling sine waves of the form

$$\mathcal{E}_y = \mathcal{E}_0 \sin 2\pi\nu\left(t - \frac{x}{v}\right) \tag{9}$$

or, similarly for H_z,

$$H_z = H_0 \sin 2\pi\nu\left(t - \frac{x}{v}\right) \tag{10}$$

These represent plane waves traveling along the positive x-axis and have a wave length $\lambda = v/\nu$, the surfaces of constant phase being given by the equation $x = constant$ at a definite instant of time. For example, let us consider one of these y-z planes, as shown in Fig. 151, in which, at a given instant of time, $\mathcal{E}$ and H have their maximum values $\mathcal{E}_0$ and H_0. We have the same values of $\mathcal{E}$ and H at every point in this plane, and this represents a "crest" of the wave. The plane containing these maximum

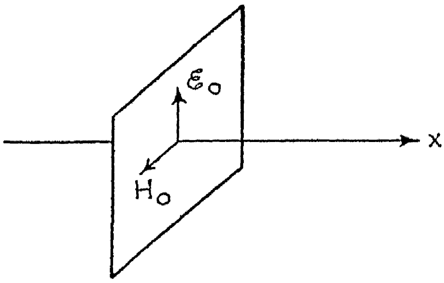

FIG. 151.

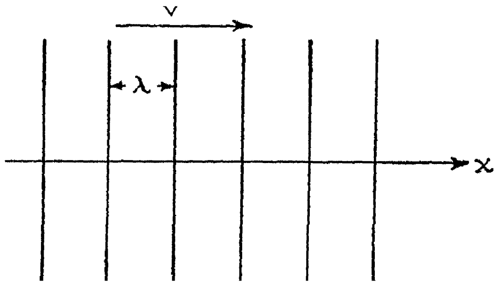

FIG. 152.

values of $\mathcal{E}$ and H moves to the right with the phase velocity v as is evident from the form of Eqs. (9) and (10). Thus we can schematically represent a traveling plane wave by a figure such as Fig. 152, in which the vertical lines represent the intersection of the crests with the plane of the paper.

In Chap. IX we derived the relation between the amplitudes $\mathcal{E}_0$ and H_0 which must exist in plane waves, and, since this relation is fundamental for our later considerations, we repeat the argument here for the case of dielectric media. Equations (9) and (10) can represent the electric and magnetic vectors of the same wave only if Eqs. (3) and (4) are satisfied. This imposes a restriction on the relative values of $\mathcal{E}$ and H. From Eq. (9) we find

$$\frac{\partial \mathcal{E}_y}{\partial x} = -\frac{2\pi\nu}{v}\mathcal{E}_0 \cos 2\pi\nu\left(t - \frac{x}{v}\right)$$

and from Eq. (10)

$$\frac{\partial H_z}{\partial t} = 2\pi\nu H_0 \cos 2\pi\nu\left(t - \frac{x}{v}\right)$$

Now Eq. (3) demands that the first of these expressions be equal and opposite in sign to the product of the second and μ/c. This yields

$$\frac{\mathcal{E}}{v} = \frac{\mu}{c} H$$

and, using the relation $v = c/\sqrt{\epsilon\mu}$, this can be written as

$$\sqrt{\epsilon}\mathcal{E} = \sqrt{\mu} H \tag{11}$$

For nonmagnetic media, Eq. (11) takes the more convenient form,

$$n\mathcal{E} = H \tag{12}$$

since $n = \sqrt{\kappa} = \sqrt{\epsilon}$ and $\mu = \mu_0 = 1$. It is left as an exercise for the student to show that this relation can also be obtained by using Eq. (4) instead of Eq. (3).

We shall need expressions for plane waves traveling in an arbitrary direction, not only along the x-axis, and we must investigate the form taken by Eq. (9), for example, for this case. First, we note that in Eq. (9) the planes of constant phase are determined by the equations $x =$ *constant*. Hence, in the expression for a plane wave traveling in an arbitrary direction, we must replace x by an expression such that, when placed equal to a constant, it yields the equations of the constant-phase surfaces, *i.e.*, planes whose normals make arbitrary angles with the x-, y-, and z-axes. Thus we must investigate the general expression for the equation of a plane. Let AA be the intersection of such a plane with the plane of the page, k a *unit* vector the direction of which is perpendicular to the plane, and r the radius vector from the origin to any point P in the plane (Fig. 153). It is clear that the projection of the radius vector r along the direction k (normal to the plane) has the same value (OQ) no matter where the point P lies in the plane. Thus the plane is the locus of all points, the radius vectors of which have the same projection OQ along the normal. Hence the equation of the plane can be written in the convenient vector form

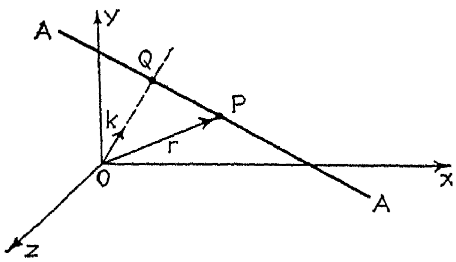

FIG. 153.

$$\vec{r} \cdot \vec{k} = OQ = \text{constant} \tag{13}$$

Now let the components of the unit vector k along the x, y, and z axes be f, g, and h, respectively. The components of $\vec{r}$ are x, y, and z (the coordinates of the point P). From the rules for forming the scalar product of two vectors, we have

$$\vec{r} \cdot \vec{k} = fx + gy + hz$$

f, g, and h are the cosines of the angles which the normal to the plane $(\vec{k})$ make with the positive x-, y-, and z-axes, respectively. It then follows that the equation of a plane in Cartesian coordinates is

$$fx + gy + hz = \text{constant} \tag{14}$$

The equation of a plane wave (of $\vec{\mathcal{E}}$) traveling in the direction $\vec{k}$ normal to the planes of constant phase can now be written. It is

$$\vec{\mathcal{E}} = \vec{\mathcal{E}}_0 \sin 2\pi\nu\left(t - \frac{fx + gy + hz}{v}\right) \tag{15}$$

where the vectors $\vec{\mathcal{E}}$ and $\vec{\mathcal{E}}_0$ are perpendicular to the direction of propagation, as they must be for a transverse wave. An exactly similar expression can be written for $\vec{H}$. Utilizing Eq. (13), we can write Eq. (15) in the more concise fashion

$$\vec{\mathcal{E}} = \vec{\mathcal{E}}_0 \sin 2\pi\nu\left(t - \frac{\vec{k} \cdot \vec{r}}{v}\right) \tag{16}$$

Note that, if $\vec{k}$ points along the positive x-axis, $f = 1$, $g = h = 0$, and Eq. (15) reduces to the form of Eq. (9), as it must.

62. Reflection and Refraction of Plane Waves.—Suppose a plane electromagnetic wave traveling with a velocity v_1 in a dielectric medium impinges on a boundary surface separating this medium from a second dielectric in which the velocity of electromagnetic waves is v_2. In accordance with the general boundary conditions developed in Chaps. XI and XII, waves will be set up in both bodies. The normal components of B and D and the tangential components of $\mathcal{E}$ and H must be continuous at the boundary surface at all points of the latter and for all values of t.

In general, it will not be possible to satisfy these conditions by postulating only a wave traveling in the second medium, but one must also require that a reflected wave be set up in the first medium. Application of the boundary conditions then yields the relations which must exist among the amplitudes, frequency, and directions of propagation of these various waves.

For the sake of simplicity let us consider a plane boundary, which we choose as the plane $x = 0$, the x-axis normal to this plane; and let the plane determined by the x-axis and the Poynting vector of the incident wave be the x-y plane* as shown in Fig. 154. In this figure are shown the directions of propagation of the incident, reflected and refracted waves, denoting the angles between these directions and the x-axis by i, r', and r, respectively. Since, for the incident wave, we have the relations

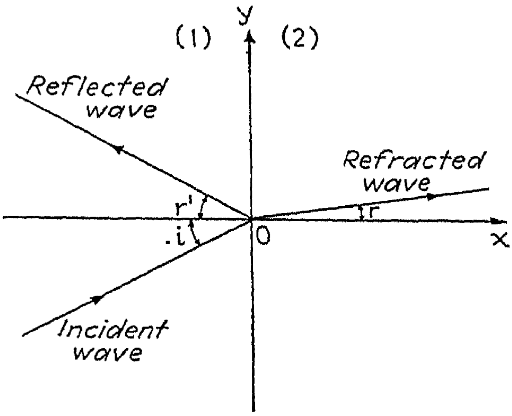

FIG. 154.

$$f = \cos i; \qquad g = \sin i; \qquad h = 0$$

the electric vector of this wave may be written in accordance with Eq. (15) as

$$\overrightarrow{\mathcal{E}_1} = \overrightarrow{\mathcal{E}_{01}} \sin 2\pi\nu_1\left(t - \frac{x \cos i + y \sin i}{}\right) \tag{17}$$

where ν_1 and v_1 are the frequency and the velocity of the wave in medium 1.

Now consider the refracted or transmitted wave. The wave normal $\overrightarrow{k}$ of this wave has direction cosines

$$f = \cos r; \qquad g = \sin r; \qquad h = 0$$

as is evident from Fig. 154. Hence the electric vector of this wave may be written in the form

$$\overrightarrow{\mathcal{E}_2} = \overrightarrow{\mathcal{E}_{02}} \sin 2\pi\nu_2\left(t - \frac{x \cos r + y \sin r}{v_2}\right) \tag{18}$$

where ν_2 and v_2 are the frequency and velocity in the second

* This plane is known as the plane of incidence.

medium. Now the boundary conditions, *e.g.*, the continuity of the tangential components of $\mathcal{E}$, must hold at all points of the boundary surface, *i.e.*, for all values of y and z when $x = 0$, and also at all instants of the time t. Comparing Eqs. (17) and (18), in which we place $x = 0$, we see that this can be true only if

$$\nu_1 = \nu_2$$

and if

$$\frac{\sin i}{v_1} = \frac{\sin r}{v_2}$$

no matter how we choose the amplitudes $\mathcal{E}_{01}$ and $\mathcal{E}_{02}$. From the first of these conditions we see that the frequencies of the waves must be the same in both media; hence the wave lengths are different. The second relation fixes the direction of propagation of the refracted wave if that of the incident wave is given. Using the relations $v_1 = c/n_1$ and $v_2 = c/n_2$, n_1 and n_2 being the indices of refraction of the two media, this relation takes the form

$$\frac{\sin i}{\sin r} = \frac{n_2}{n_1} \tag{19}$$

The ratio of the sine of the angle of incidence (i) *to the sine of the angle of refraction* (r) *equals the ratio of the indices of refraction.* This is the well-known law of refraction in optics, and is called *Snell's law.*

Finally, let us consider the reflected wave. Its wave normal has the direction cosines

$$f = \cos(\pi - r') = -\cos r'; \qquad g = \sin(\pi - r') = \sin r'; \qquad h = 0$$

remembering that the angles are measured with respect to the *positive* x-axis. Hence Eq. (15) for the electric vector of this reflected wave takes the form

$$\overrightarrow{\mathcal{E}_1'} = \overrightarrow{\mathcal{E}_{01}'} \sin 2\pi\nu_1\left(t + \frac{x\cos r' - y\sin r'}{v_1}\right) \tag{20}$$

and we can repeat the arguments of the previous paragraph. Since the boundary conditions must hold at all points on the surface $x = 0$, we must have, comparing Eqs. (17) and (20),

$$\sin i \qquad \sin r'$$

or

$$i = r' \tag{21}$$

The angle of incidence equals the angle of reflection, another familiar law of elementary optics.

Let us now examine some of the consequences of Snell's law, Eq. (19). There are two cases to consider: (1) the case for which $n_2 > n_1$ and (2) the case for which $n_1 > n_2$. In the first case, speaking of optical waves, we say that the wave travels from an optically "rarer" medium to an optically "denser" medium, and conversely for the second case. Since $\sin i$ can take on all values from zero to unity, $\sin r$ [which is equal to $(n_1/n_2) \sin i$] takes on corresponding values lying between zero and n_1/n_2. Now in the case for which $n_1/n_2 < 1$, this corresponds to a real angle of refraction for every angle of incidence. On the other hand, for waves traveling from an optically denser medium into a rarer one, $n_1/n_2 > 1$, and in this case refraction cannot take place for all angles of incidence. If the angle of incidence is smaller than $\sin^{-1}(n_2/n_1)$, then $\sin r$ has a value between zero and unity, and a refracted wave exists. For angles of incidence larger than this value, *i.e.*, if $\sin i > n_2/n_1$, the angle of refraction becomes imaginary, and there is no refracted wave, only a reflected one. For such a case one speaks of *total reflection.*

When one applies the boundary conditions to the waves described by Eqs. (17), (18), and (20), and to the corresponding expressions for H, the resulting equations fix the relative values of the amplitudes and hence the intensities of the incident, reflected, and refracted waves. The laws of reflection and refraction, embodied in Eqs. (19) and (21), must hold in any case, provided these waves are present. These laws yield information as to the relative directions of the waves but leave the question of relative intensities untouched. We can, however, answer the last question with the help of the method indicated above and thus see that our fundamental electromagnetic equations embody, not only the laws of so-called geometrical optics, but also those of physical optics.

63. Intensity Relations for Normal Incidence.—In this section we shall carry through the calculations for intensities in the special case of normal incidence, *i.e.*, when the wave normal of the incident wave coincides with that of the boundary (Fig. 155). Let $\mathcal{E}_1$ be the electric vector of the incident wave and $\mathcal{E}_1'$

and $\mathcal{E}_2$ those of the reflected and refracted (transmitted) waves, respectively. At a given instant of time the magnetic vectors, H_1, H_1', and H_2 all point into the plane of the paper, and the corresponding directions of the electric vectors are shown. In the reflected wave $\mathcal{E}_1'$ must be opposite to $\mathcal{E}_1$, so that the Poynting vector S represents a wave traveling along the negative x-axis. One must have either $\mathcal{E}$ or H reversed in phase for the reflected wave, and we choose $\mathcal{E}$ arbitrarily as the one which is changed. Our final equations will answer uniquely the question as to which vector suffers a 180° phase change on reflection.

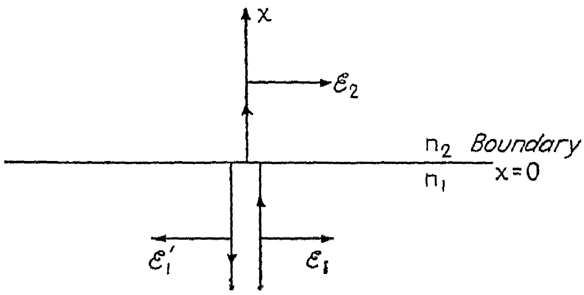

FIG. 155.

The following conditions must be satisfied at the boundary $x = 0$:

1. The normal components of D must be continuous at the boundary.
2. The normal components of B must be continuous at the boundary.
3. The tangential components of $\mathcal{E}$ must be continuous at the boundary.
4. The tangential components of H must be continuous at the boundary.

In our special case of normal incidence the first two conditions are obviously satisfied, since all the vectors are parallel to the boundary surface. Condition 3 yields

$$\mathcal{E}_1 - \mathcal{E}_1' = \mathcal{E}_2 \tag{22}$$

and condition 4 gives

$$H_1 + H_1' = H \tag{23}$$

In these equations the $\mathcal{E}$'s and H's denote the magnitudes of the vectors for $x = 0$ at any instant of time. Since the vectors $\mathcal{E}_1$, $\mathcal{E}_1'$, and $\mathcal{E}_2$ are given by

$$\mathcal{E}_1 = \mathcal{E}_0 \sin 2\pi\nu\left(t - \frac{n_1 x}{c}\right)$$

$$\mathcal{E}_1' = \mathcal{E}_0' \sin 2\pi\nu\left(t + \frac{n_1 x}{c}\right)$$

$$\mathcal{E}_2 = \mathcal{E}_{02} \sin$$

SI

in accordance with the general Eqs. (17), (18), and (20), we see that only at the boundary, $x = 0$, can Eq. (22) be satisfied for all values of t. Now we make use of the relation between the magnitudes of $\mathcal{E}$ and H in a plane wave and have from Eq. (12)

$$H_1 = n_1\mathcal{E}_1; \qquad H_1' = n_1\mathcal{E}_1'; \qquad H_2 =$$

Equation (23) can then be written in the form

$$\mathcal{E}_1 + \mathcal{E}_1' = \frac{n_2}{n_1}\mathcal{E}_2 \tag{24}$$

Equations (22) and (24) show the necessity of assuming the existence of *both* reflected and refracted waves. Were either assumed missing, we could not simultaneously satisfy both these equations. From these equations there follows,

$$\mathcal{E}_1' = \frac{n_2 - n_1}{n_1 + n_2}\mathcal{E}_1 \tag{25}$$

giving the electric vector of the reflected wave in terms of that of the incident wave. For the transmitted wave one finds

$$\mathcal{E}_2 = \frac{2n_1}{n_1 + n_2}\mathcal{E}_1 \tag{26}$$

Equation (25) now shows us that, if $n_2 > n_1$, the wave impinging on an optically denser medium, the electric vector at the surface suffers a phase change of 180° upon reflection, and the magnetic vector undergoes no phase change, as we assumed. On the other hand, if $n_2 < n_1$, as one would have when a beam of light travels from glass to air, the electric vector of the reflected wave is in phase with that of the incident wave at the surface, whereas the magnetic vectors are 180° out of phase with each other.

One is generally more interested in the intensity relations than in the amplitude relations. The energy incident per unit area on the boundary surface per unit time is given by the Poynting vector of the incident wave and is

$$S_1 = \frac{c}{4\pi}(\mathcal{E}_1 \times H_1) = \frac{n_1 c}{4\pi}\mathcal{E}_1^2$$

since $H_1 = n_1\mathcal{E}_1$.

The reflected intensity is

$$S_1' = \frac{c}{4\pi}(\mathcal{E}_1' \times H_1') = \frac{n_1 c}{4\pi}\mathcal{E}_1'^2$$

so that the ratio of the reflected to incident intensities is given by

$$R = \frac{S_1'}{S_1} = \frac{\mathcal{E}_1'^2}{\mathcal{E}_1^2}$$

R is known as the reflecting power of the surface. Using Eq. (25), we then have for R

$$R = \left(\frac{n_2 - n_1}{n_1 + n_2}\right)^2 \tag{27}$$

as the reflecting power at normal incidence. The reflecting power is always less than unity and approaches this value as n_2 becomes large compared to n_1, or vice versa. For a glass-air boundary, the glass having an index of refraction of about 1.5 and setting the index of refraction of air equal to 1, we obtain as the reflecting power of the glass surface

$$R = \left(\frac{0.5}{2.5}\right)^2 = \frac{1}{25} = 0.04$$

Thus about 4 per cent of the intensity of a light beam falling normally on a glass surface is reflected.

The calculation of the intensity relations at boundaries for the case of an arbitrary angle of incidence follows the same scheme as for normal incidence. It is more involved however, since the reflecting power depends on the polarization of the incident light, *i.e.*, on whether the electric vector oscillates in, or at right angles to, the plane of incidence, and we shall not carry it through.

The propagation of electromagnetic disturbances in conducting media, such as metals, is a much more complicated phenomenon than in dielectrics, so that we must content ourselves in this treatment with a few qualitative remarks. If we consider metals and assume the validity of Ohm's law, the Ampère circuital law must be extended from the special form employed in this chapter to take into account conduction currents. Thus it would take the form

$$\oint H_s\,ds = \frac{1}{c}\int \frac{\partial D_n}{\partial t}dS + \frac{4\pi\sigma}{c}\int \mathcal{E}_n\,dS$$

The effect of the last term on the right would be to modify our equations so that, even in the simple case of an electric field with but one component, say $\mathcal{E}_y$, which depends only on x and t, $\mathcal{E}_y$ does not satisfy the wave equation but a more complicated equation. It is still true that the disturbances will be transverse, and they will to some extent resemble ordinary transverse waves. The essential differences for plane electromagnetic waves in metals as compared to those in dielectrics may be summarized as follows:

1. The amplitude of the vector ($\mathcal{E}$ or H) decreases exponentially as x increases. Thus the Poynting vector decreases as the wave travels, and the rate of decrease of this vector is a measure of the joule heating produced in the metal. We say that we have a space-damped wave.

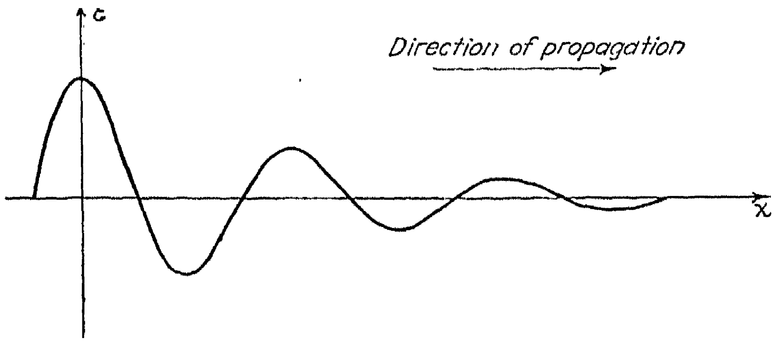

FIG. 156.

2. The velocity of propagation depends on the frequency, even if ϵ and σ are assumed independent of frequency.

3. The magnetic and electric vectors are not in phase with each other, as they are in the case of plane waves in dielectrics.

In Fig. 156 is shown the variation of the amplitude of the electric vector with distance in the direction of propagation of an electromagnetic wave in a metal.

Problems

1. Assuming linearly polarized plane waves traveling in a nonconducting medium of the form of Eqs. (9) and (10), show, by using Eq. (4) of the text, that the electric and magnetic vectors are related by the equation

$$\sqrt{\epsilon}\mathcal{E} = \sqrt{\mu}H$$

2. Given a cube of edge a, the edges lying along the x-, y-, and z-axes, and the origin O at one corner of the cube.

a. What is the angle between the body diagonal from the origin O to the opposite corner P and an edge of the cube?

b. Find the equation of a plane perpendicular to this body diagonal which contains the point P.

c. At what points does this plane intersect the x-, y-, and z-axes?

3. A plane wave of light is incident on one side of a glass plate of thickness d.

a. Show that the plane wave emerging from the other side of the plate has the same direction of propagation as the incident wave.

b. Consider a given normal of the incident wave. Prove that, as the wave passes through the glass, this normal undergoes a lateral displacement given by

$$\frac{d \sin (i - r)}{\cos r}$$

where i and r are the angles of incidence and refraction at the first glass surface.

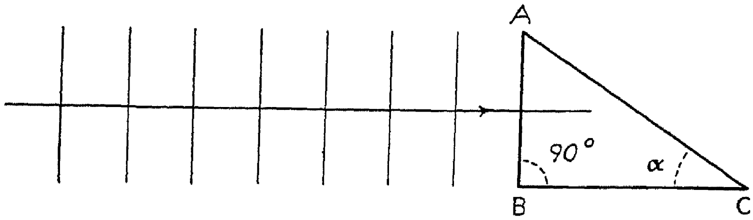

FIG. 157.

4. A plane of light falls on the face AB of a glass prism at normal incidence, as shown in Fig. 157. The index of refraction of the glass is 1.50. Find the smallest or largest value of the angle α such that the wave will be totally reflected at the surface AC:

a. If the prism is surrounded by air.

b. If the prism is surrounded by a liquid of refractive index 1.40.

5. Liquid of refractive index 1.63 stands at a height of 2.00 cm. in a flat-bottomed glass vessel. The refractive index of the glass is 1.50. Show whether or not a plane wave of light incident on the top surface of the liquid can be totally reflected at the bottom surface.

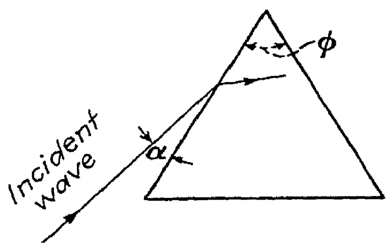

FIG. 158.

6. Find the largest value of the angle ϕ of the glass prism in Fig. 158 such that a light wave incident as shown will pass through the prism:

a. When the prism is surrounded by air.

b. When the prism is surrounded by water.

The angle α is small enough so that $\cos \alpha$ is practically equal to unity. The index of refraction of the glass is 1.55 and that of water is 1.33.

7. A plane light wave passes normally through a glass plate with plane parallel faces.

a. Compute the ratio of the intensity of the transmitted to that of the incident wave taking into account only one reflection at each of the glass faces.

b. Set up a general expression for the above ratio, taking into account all the internal reflections inside the glass.

8. A plane light wave is normally incident on the liquid surface in Prob. 5. If the electric vector of the incident wave has a maximum value of 10^{-3} volt/meter, compute the intensity of the wave transmitted through the glass vessel. (Consider only one reflection at each interface.)

9. Two dielectrics of indexes of refraction n_1 and n_2 are separated by a plane boundary. A plane wave in the medium of index n_1 falls on the boundary at normal incidence. Compute expressions for the Poynting vector for the incident, reflected, and transmitted waves, and show that the energy incident on the boundary per unit time is equal to the sum of the energies carried away from the boundary per unit time by the reflected and transmitted waves.

10. *a.* Show, for the case of an electromagnetic disturbance traveling along the x-axis in a metal, that Eqs. (3) and (4) of the text become

$$\frac{\partial \mathcal{E}_y}{\partial x} = -\frac{1}{c}\frac{\partial B_z}{\partial t}$$

$$\frac{\partial H_z}{\partial x} = -\frac{1}{c}\frac{\partial D_y}{\partial t} - \frac{4\pi\sigma}{c}\mathcal{E}_y$$

where σ is the conductivity of the metal.

b. From these equations show that $\mathcal{E}_y$ and H_z both satisfy equations of the form

$$\frac{\partial^2 \mathcal{E}_y}{\partial x^2} = \frac{\epsilon\mu}{c^2}\frac{\partial^2 \mathcal{E}_y}{\partial t^2} + \frac{4\pi\sigma\mu}{c^2}\frac{\partial \mathcal{E}_y}{\partial t}$$

11. Following the argument of Sec. 44, Chap. VIII, show that in a metal the rate of decrease of field energy in a volume element $dx\,dy\,dz$ is equal to the net rate at which energy flows out of this element (computed from the Poynting vector) plus the rate of joule heating in this element. Use the equations of Prob. 10, part *a.*

CHAPTER XIV

GEOMETRICAL OPTICS AND SIMPLE OPTICAL INSTRUMENTS

In Chap. XIII we have seen how the fundamental laws of electromagnetism led to the laws of the reflection and refraction of plane waves at a boundary separating two dielectric media. By far the most important practical application of these laws is to the case of the reflection and refraction of light waves, electromagnetic waves of wave lengths ranging from about 4 $\times 10^{-5}$ cm. to about 7×10^{-5} cm. in air, at the surfaces of mirrors and lenses. In this chapter we shall concern ourselves specifically with this type of problem. The general problem of following the propagation of electromagnetic waves is far too complicated to allow a complete analysis in this book. We may, however, make a few remarks concerning some of the simpler aspects of the general method, which is embodied in a principle known as *Huygens' principle.* Suppose that we know the shape of one of the constant-phase surfaces of a wave, *e.g.*, one of the crests of the wave, at some instant of time. We can find the shape of this wave surface at a later time Δt by considering each point on the original wave surface as a source of secondary spherical waves which diverge from these points. If one constructs spheres of radii equal to $v\,\Delta t$, v being the phase velocity of the waves, with centers at the various points on the initial wave surface, the envelope of these spherical wavelets then yields the shape of this wave surface at a time Δt later. Thus in Fig. 159 there is shown the trace of the initial wave surface AA and its trace at a time Δt later. This geometrical construction (Huygens' construction) is only part of the story, however, and offers no advantage over the simpler method of *rays* which we shall discuss shortly. To complete the analysis, we must know how

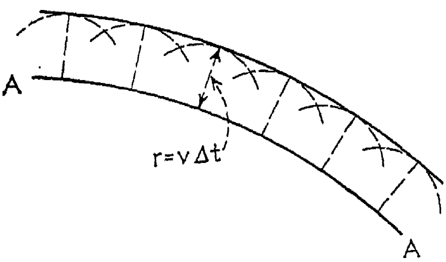

Fig. 159.

the amplitudes of these secondary wavelets vary with direction, and it turns out that this variation of amplitude with direction of propagation is rather complicated. It is just the solution of this part of the problem which is prohibitively difficult.

In our study of plane waves we have seen that the constant-phase surfaces could be described by constructing the normals to these surfaces (the wave normals or *rays*), and one can follow the motion of these surfaces by moving along the directions of these rays. This mode of description is evidently possible for waves other than plane waves. For example, in the case of spherical waves in a homogeneous medium, the rays consist of straight lines radiating in all directions from a common point, and the surfaces of constant phase are concentric spherical surfaces. If the Poynting vector is directed outward from the center, one speaks of a *diverging* wave, if toward the center, of a *converging* wave. One is often interested in following the motion of a limited portion of a wave surface, and one can construct a bundle or *pencil* of rays through this portion of the surface. Such a pencil of rays is called a *beam*, and for plane waves the beam consists of a parallel bundle of rays. For a spherical portion of a wave surface the rays diverge from (or converge on) a point F, known as the *focal point* of the pencil, as indicated in Fig. 160. Such a pencil is called *stigmatic*. The pencil of rays from a portion of a wave surface which has different radii of curvature in two mutually orthogonal directions (such as one shaped like a blowout patch) form an *astigmatic* pencil and do not pass through a common point. In the case of the propagation of waves in a homogeneous medium the rays always are straight lines, and the wave surfaces do not change shape as the wave propagates. If, however, the velocity varies from point to point, the rays will be curved lines, and the wave surfaces will not maintain an unaltered shape.

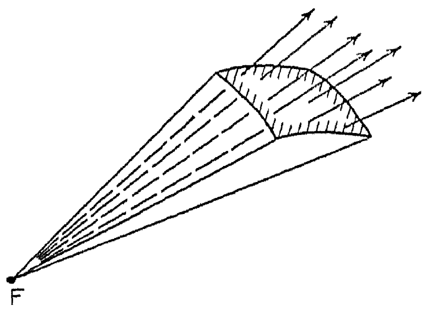

FIG. 160.

Thus far our remarks are valid for waves of any wave length, and the advantage gained by describing wave motion in terms of rays becomes evident if we consider what happens when we try to form a narrow beam from a plane wave by allow-

ing the latter to fall on a screen, in which there is a hole, placed perpendicular to the direction of propagation. The waves emerging from the hole will, in general, *not* form a section of a plane wave with a parallel bundle of rays, but will spread out more or less in all directions. As we shall see later, *if the wave length of the waves is very small compared to the linear dimensions of the aperture,* this spreading effect, or *diffraction,* as it is called, becomes extremely small, and, just in the case of light waves, the wave length is very small compared with the dimensions of ordinary objects. In this chapter we shall neglect diffraction effects and treat the bundle of light rays emerging from the aperture as strictly parallel. Similarly, we shall assume that an obstacle placed in the path of a beam of light casts a sharply defined geometrical shadow. The laws of optics and optical systems, to the approximation in which one can neglect typical wave effects such as diffraction and interference, comprise the subject of *geometrical optics* and in this chapter we shall concern ourselves with this study.

64. Fermat's Principle.—The calculation of the path of rays of light may be effected with the help of a general principle due to Fermat, bearing his name. The principle states that the path of a ray between any two points P_1 and P_2 will be such that the time required for light to traverse the path will be a minimum. This principle is valid for media with varying indices of refraction or for the case in which the light ray passes from one medium to a second with a different index of refraction. Let us formulate this principle quantitatively. Consider an arbitrary path connecting two points P_1 and P_2 as shown in Fig. 161, and let us compute the time necessary for light to pass along this path from P_1 to P_2. The time needed to traverse a portion of length ds is ds/v, where v is the velocity of light at the point where ds is situated. The total time is then

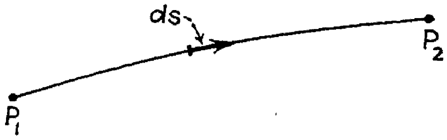

FIG. 161.

$$t = \int_{P_1}^{P_2} \frac{ds}{v} = \frac{1}{c}\int_{P_1}^{P_2} n\, ds \tag{1}$$

where we have set $v = c/n$, n the index of refraction of the medium. The integral $\int_{P_1}^{P_2} n\, ds$ is called the *optical length* of the path from P_1 to P_2. In Eq. (1) the integral is a line integral and will have different values for different paths connecting the end points P_1 and P_2. Now suppose we choose a path only slightly displaced from the one shown in Fig. 161. The difference of time of traversal for the two paths will, in general, be of the same order of

magnitude as that of a quantity determining the displacement of the paths. Fermat's principle states that, if we have the correct path, the difference of time of traversal of it and a neighboring path will be of the order of magnitude of the square or higher powers of the quantity measuring the displacement of the two paths. This is analogous to the statement in ordinary calculus that a function $f(x)$ has a minimum (or maximum) at the point $x = a$. Suppose we wish the value of this function at a point x close to $x = a$. Then by Taylor's theorem we can write

$$f(x) = f(a) + \left(\frac{df}{dx}\right)_{x=a}(x - a) + \frac{1}{2}\left(\frac{d^2f}{dx^2}\right)_{x=a}(x - a)^2 + \cdots$$

and, if the function has a minimum at $x = a$, we have $df/dx = 0$ for $x = a$, and find:

$$f(x) - f(a) = \frac{1}{2}\left(\frac{d^2f}{dx^2}\right)_{x=a}(x - a)^2 + \cdots$$

so that the difference in the value of the function at x and at a is of the order of magnitude of $(x - a)^2$. On the other hand, if there is no minimum at $x = a$, $f(x) - f(a)$ is of the order of magnitude of $(x - a)$.

Let us see how this principle works for the simple case of a light ray in a homogeneous medium. Then the time of traversal is simply $\frac{1}{v}\int_{P_1}^{P_2} ds$, and this is a minimum for the straight-line path connecting P_1 and P_2. Now consider a slightly varied path such as P_1AP_2 (Fig. 162). The length of the path P_1AP_2 is

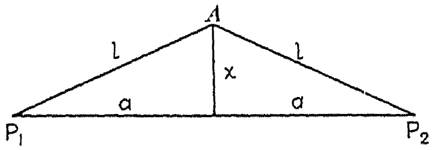

FIG. 162.

$$2l = 2\sqrt{a^2 + x^2}$$

and the time of traversal is

$$t' = \frac{2l}{v} = \frac{2a}{v}\left[1 + \left(\frac{x}{a}\right)^2\right]^{\frac{1}{2}}$$

The traversal time for the correct path is

$$t = \frac{2a}{v}$$

so that the difference is

$$t' - t = \frac{2a}{v}\left[1 - \sqrt{1 + \left(\frac{x}{a}\right)^2}\,\right]$$

Using the binomial theorem, we can expand $\sqrt{1 + \left(\frac{x}{a}\right)^2}$ and find

$$t' - t = \frac{a}{v}\left(\frac{x^2}{a^2} + \cdots\right)$$

which is of the order of magnitude of the square of the small quantity x. Thus the straight line is the correct path. Let us carry out the same sort of calculation for two neighboring incorrect paths as shown in Fig. 163, where x is a small quantity. The difference of time is clearly

$$t' - t = \frac{2}{v}(l' - l)$$

Now $(l' - l) = AB$ and AB is very nearly equally $x \sin \theta$, so that

$$t - t' = \frac{2x \sin \theta}{v}$$

which is now of the order of magnitude of x, provided $\sin \theta \neq 0$.

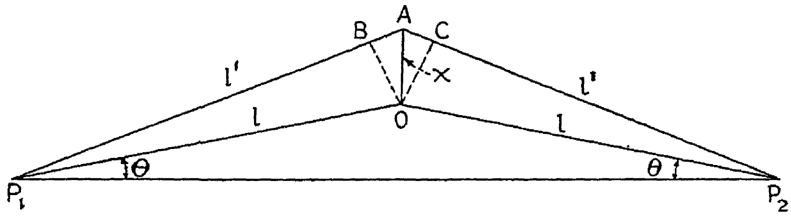

FIG. 163.

In applying Fermat's principle to find the correct path of a light ray, we proceed as follows: Compute the times needed for light to pass from one point to another along any two neighboring paths. Form the difference of these times, retaining only terms of the *first* order in small quantities. Setting these equal to zero then gives the correct path.

As an example of this procedure we shall derive the law of refraction for a ray of light passing across a boundary which separates two media of refractive indices n_1 and n_2. In Fig. 164 we show two rays starting from a point P_1 in medium 1 and reaching a point P_2 in medium 2. The ray P_1AP_2 travels a longer distance BA in medium 1, but the ray POP_2 travels a longer distance OC in medium 2. Thus the difference of time Δt is

$$\Delta t = \frac{BA}{v_1} - \frac{OC}{v_2}$$

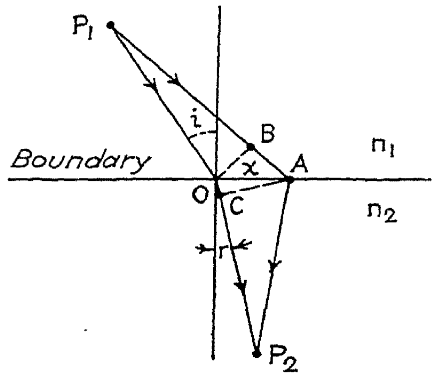

FIG. 164.

The distance BA is equal to $x \sin i$ and OC is $x \sin r$. These expressions are approximate but differ from the exact expressions by terms which involve higher powers of x than the first, and hence these suffice for our purpose. Thus we find

$$\Delta t = x\left(\frac{\sin i}{v_1} - \frac{\sin r}{v_2}\right) + \cdots$$

where the dots indicate terms involving x^2 and higher powers. Thus the correct path occurs when

$$\frac{\sin i}{v_1} - \frac{\sin r}{v_2} = 0$$

or

$$n_1 \sin i = n_2 \sin r \tag{2}$$

and this is the law of refraction which we derived in Chap. XIII. Further examples are left to the problems.

65. Reflection of Light.—We have derived the law of reflection, angle of incidence equals angle of reflection, for the case of plane waves reflected from a plane surface. It is justifiable, however, to use it for other types of waves, *e.g.*, spherical waves, reflected from curved surfaces. We can see this as follows: Consider an infinitesimal element of area on the wave surface of an arbitrary wave. The normal to this elementary area gives the direction of the ray at the point where the element is located, and, since we are dealing with an infinitesimal area, it may be considered as plane. Now, if this wave impinges on a curved surface, our elementary area will come in contact with an infinitesimal section of the reflecting surface, and the latter may be considered as plane. Thus we can apply the law of reflection to this particular ray, and we must simply take into account the fact that different rays of the same beam have different directions and that the normal direction to the reflecting surface varies from point to point.

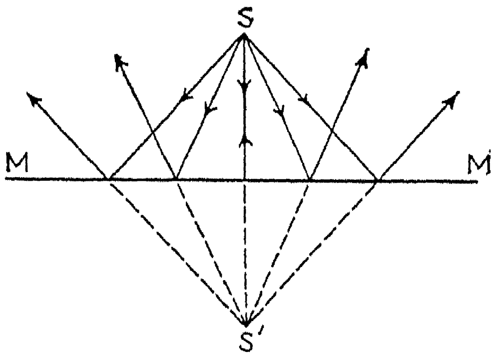

FIG. 165.

For example, consider a spherical wave diverging from a point source of light S and incident on a plane mirror. In Fig. 165 we show a diverging beam of rays from S incident on the mirror MM and the reflected beam of rays which appear to be diverging from a source S' in back of the mirror. This focal point S' of the reflected rays is called the image of the source S, and we say that it is a *virtual* image since the rays do not actually pass through this point. There are cases where a pencil of rays actually do pass through a definite point in a medium and then diverge from that point. In such cases one speaks of a *real* image. One can

readily show that an object of finite size is imaged in a plane mirror in such a manner that each point of the image is just as far back of the surface as the corresponding point of the object is in front of it and that the linear dimensions of object and image are identical. Thus the linear magnification, the ratio of linear dimension of image to object, is unity for a plane mirror.

An important case of reflection is that in which the reflecting surface is a portion of a spherical surface, a so-called spherical mirror. In the elementary discussion of spherical mirrors one learns that a bundle of rays parallel to the mirror axis (the normal to the mid-point of the mirror) is brought to a focus at a point halfway between the center of curvature of the mirror and the intersection of the mirror surface and the axis (the so-called vertex). This statement is true only if the rays lie very close to the mirror axis, so that in practice it can be applied only to mirrors of small aperture, *i.e.*, when the mirror forms only a very small part of the whole spherical surface. Let us examine the general case. In Fig. 166 are shown the mirror and an incident ray parallel to the axis, which we call the x-axis. The vertex of the mirror is taken as an origin, and we wish to find an expression for the distance x, the point Q being the point where the reflected ray crosses the mirror axis. In the triangle CPQ we have immediately from the law of sines

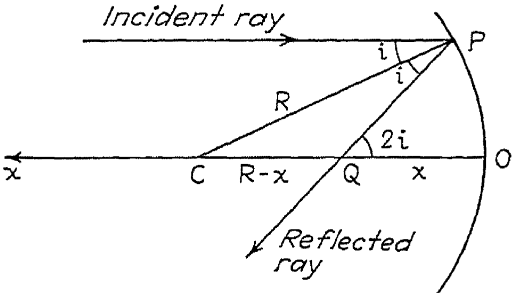

FIG. 166.

$$\frac{R - x}{R} = \frac{\sin i}{\sin (\pi - 2i)} = \frac{\sin i}{\sin 2i} = \frac{1}{2 \cos i}$$

and solving for x, we find

$$x = R\left(1 - \frac{\;}{2 \cos i}\right) \tag{3}$$

Thus we see that the point at which a reflected ray crosses the axis depends on the angle of incidence of the ray, the value of x decreasing as the angle of incidence increases. Only in the case of angles small enough so that we may place $\cos i = 1$ do we find the reflected rays all passing through a single point. For this case we have from Eq. (3)

$$x_0 = R\left(1 - \frac{1}{2}\right) = \frac{R}{2} \tag{4}$$

and this point x_0 is called the *principal focus* of the mirror and the distance x_0 is called the *focal length* of the mirror. The departure from sharp focusing of a bundle of paraxial rays coming from infinity by a mirror of large aperture is called *spherical aberration*. If the mirror surface is in the form of a paraboloid of revolution, objects very far from the mirror (at infinity) will be brought to a sharp focus. In astronomical mirrors one is always interested in imaging objects which are practically at infinity, so that paraboloidal mirrors are invariably used.

If we restrict ourselves to the case of spherical mirrors of such small aperture that all the rays diverging from an object (which

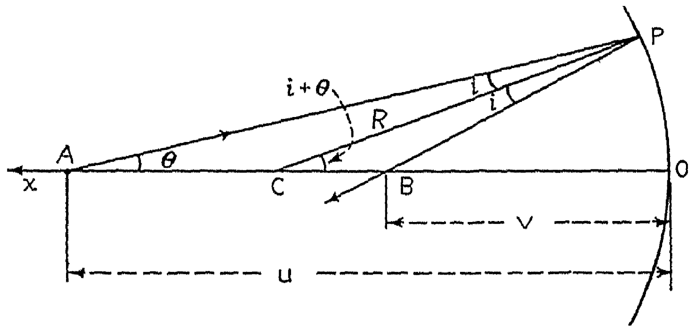

FIG. 167.

may be at a finite distance from the mirror) make very small angles with the mirror axis, then the image will be sharp. We shall compute the position of the image by considering the object as a point on the mirror axis. First we note that a ray from the object coincident with the mirror axis is reflected on itself, so that the image must lie on the mirror axis.

In Fig. 167 we show an incident ray from the object A making an angle θ with the mirror axis, and the reflected ray intersecting this axis at B. The distance OA is denoted by u and is called the object distance and $OB = v$ the image distance. In the triangle CPB we have from the law of sines

$$\frac{R - v}{R} = \frac{\sin i}{\sin\,[\pi - (2i + \theta)]} = \frac{\sin i}{\sin\,(2i + \theta)} \tag{5}$$

and using the triangle APC there follows

$$\frac{u - R}{R} = \frac{\sin i}{\sin \theta} \tag{6}$$

By eliminating the angle i from Eqs. (5) and (6), we can follow the reflected ray for every value of θ, and in general v will depend on θ. If the angles i and θ are small enough so that we may set $\sin i = i$ and $\sin \theta = \theta$, simultaneous solution of Eqs. (5) and (6) yields easily

$$\frac{1}{u} + \frac{1}{v} = \frac{2}{R}$$

which is the usual expression relating image and object distances for a spherical mirror. Essentially the same formulas hold for the case of a convex mirror, and the derivations for this are left to the problems.

In any case it is clear that one may always find the image of any object by graphical construction, tracing a bundle of rays diverging from the object and, using the fact that the angle of incidence equals the angle of reflection, then tracing the corresponding reflected rays.

66. Refraction of Light at a Spherical Surface.—The refraction of light at a spherical interface between two transparent media is fundamental in the study of optical instruments, since plane and spherical surfaces are the only ones which can be produced at sufficiently low cost for most practical purposes. In problems of this type, we have to do with incident light in a medium of refractive index n_1, let us say (the so-called *object* space), and refracted light in a medium of refractive index n_2 (the so-called *image* space). In order to minimize the chance of algebraic errors, it is essential to adopt a set of conventions for the coordinate systems to be employed and for the algebraic signs of the distances appearing in the calculations and to adhere rigidly to them. We shall adopt the following conventions:

1. Draw all figures with the light incident on the refracting surface from the left.

2. In object space measure positive object distances to the *left* along the axis of the system from an origin which, in the case of a single refracting surface, is located at the vertex of the refracting surface.

3. In image space measure positive image distances to the *right* along the axis of the system from an origin which, in the case of a single refracting surface, is located at the vertex of the refracting surface.

4. Treat radii of curvature as positive distances when the center of curvature lies to the *right* of the vertex and as negative when the center of curvature lies to the *left* of the vertex.

Consider the refraction of a pencil of rays diverging from an object point A on the axis of symmetry of a spherical refracting surface, as shown in Fig. 168.

Let AP be one of the rays intercepted by the refracting surface of radius R (positive for the case of Fig. 168), and let its direction relative to the axis be θ_1. The refracted ray crosses the axis at the point B, and the problem is to determine the position of

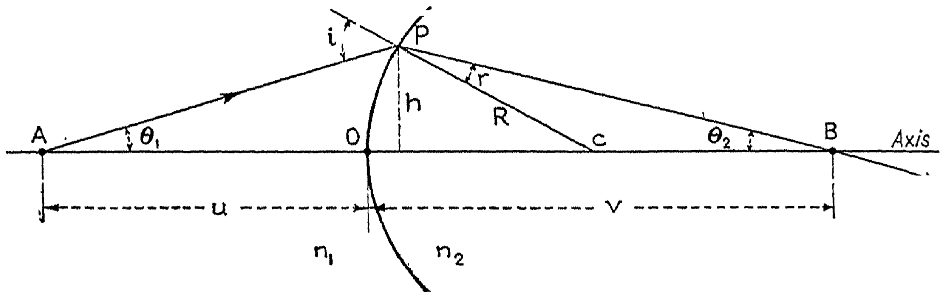

FIG. 168.

this point for an arbitrary ray coming from the object A. Applying the law of sines to the triangle APC, we have

$$\frac{u+R}{R} = \frac{\sin(\pi - i)}{\sin\theta_1} = \frac{\sin i}{\sin\theta_1}$$

and, using the triangle BPC,

$$\frac{v-R}{R} = \frac{\sin r}{\sin\theta_2} \tag{9}$$

If we divide Eq. (8) by Eq. (9) and use the law of refraction ($n_1 \sin i = n_2 \sin r$), there follows

$$\frac{u+R}{v-R} = \frac{n_2}{n_1}\frac{\sin\theta_2}{\sin\theta_1} \tag{10}$$

Finally, the angle θ_2 may be expressed in terms of the angles θ_1, i, and r by equating the sum of the angles in the triangle APB to π. If this is done, we see in general that the image distance v will be different for different angles θ_1 of the incident ray, and this gives rise to spherical aberration of the same sort as we encountered in the case of reflection.

For the special but extremely important case that all the rays from A intercepted by the refracting surface make very small

angles θ_1 with the axis (the so-called paraxial rays), we find that the image distance is the same for all rays coming from A. In this approximation the object A is focused at B. If the angles θ_1 and θ_2 are sufficiently small, we may write approximately (see Fig. 168)

$$\sin \theta_1 = \frac{h}{u}; \qquad \sin \theta_2 = \frac{h}{v}$$

so that

$$\frac{\sin \theta_2}{\sin \theta_1} = \frac{u}{v} \tag{11}$$

Substituting this value in Eq. (10), there follows

$$\frac{u + R}{R - v} = -\frac{n_2}{n_1} \cdot \frac{u}{v} \tag{12}$$

an equation from which the angle θ_1 has disappeared. This equation, although derived for rays making small angles with the axis, is still valid for large angles when the surface has been corrected for spherical aberration, and is fundamental in geometrical optics. Equation (12) is usually written in the more convenient equivalent form

$$\frac{n_1}{u} + \frac{n_2}{v} = \frac{n_2 - n_1}{R} \tag{13}$$

Although we have derived Eq. (13) for the case of a convex refracting surface (R positive), it holds equally well for a concave surface (R negative), as can be readily demonstrated. The convention as to the algebraic signs of u and v must be kept in mind when applying Eq. (13).

In optical systems one is frequently confronted with a situation in which a series of refracting surfaces are employed. The method to be employed in computing, let us say, the position of the image of an object in such a case, is to successively apply Eq. (13) to each refracting surface, treating the image formed by the first surface as the object for the second, and so on. If the image which would be formed in the presence of the first refracting surface alone lies to the left of the second refracting surface (using our conventions), it acts as a *real* object for the second surface, whether it is a real or virtual image of the original surface. If, however, the second refracting surface is so placed that a converging pencil of rays from the first surface is inter-

cepted before coming to a focus, the object distance for the second refracting surface is to be taken as negative and equal numerically to the distance from the vertex of the second refracting surface to the focal point of the pencil of rays from the first refracting surface. Such a case is shown in Fig. 169, in which $O'A$ is the object distance to be used in computing the refraction at surface 2.

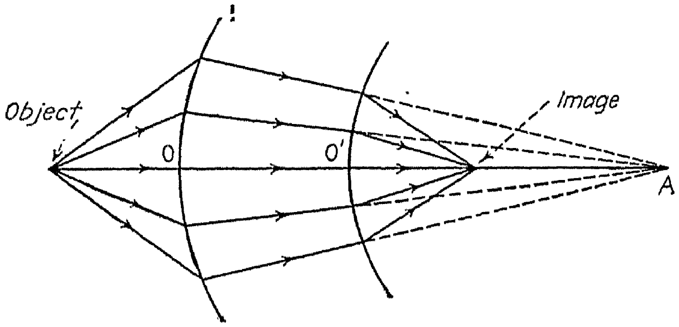

FIG. 169.

Concisely stated, we treat the image space of the nth surface as the object space of the $(n + 1)$st surface, using the appropriate indices of refraction in Eq. (13).

Consider an object of linear dimension y perpendicular to the axis of a single refracting surface, as shown in Fig. 170, and let the corresponding linear dimension of the image be y'. The linear dimension y' may be found by constructing the ray shown

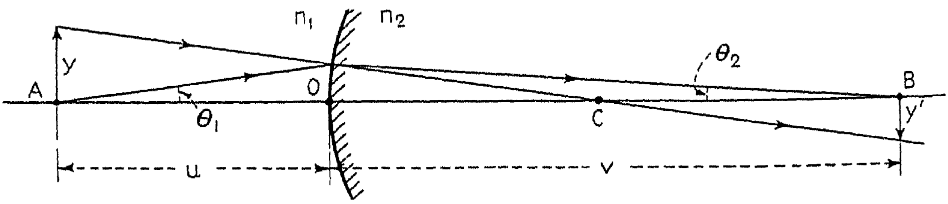

FIG. 170.

which passes through the center of curvature C of the refracting surface. This ray is not refracted as it is normally incident on the surface. We shall treat y as positive and y' as negative, corresponding to the ordinary conventions of coordinate geometry. Furthermore, we define the *linear magnification* (more precisely, the linear lateral magnification) m as the ratio of y' to y. From Fig. 170 it is evident that

$$-\frac{y'}{y} = \frac{BC}{AC} = \frac{v - R}{u + R}$$

and, using Eq. (10), we may write for the magnification

$$m = \frac{y'}{y} = -\frac{n_1}{n_2}\frac{\sin\,\theta_1}{\sin\,\theta_2} \tag{14}$$

This equation is known as "Abbe's sine condition." To the approximation of paraxial rays, we may utilize Eq. (11) instead of Eq. (10) and obtain

$$m = \frac{y'}{y} = -\frac{n_1}{n_2}\frac{v}{u} \tag{15}$$

67. Thin Lenses.—By far the most important application of the results obtained in the preceding section is to the case of lenses in air. By a lens one means a portion of glass, or some other transparent substance, which is usually bounded by plane or spherical surfaces. It is assumed that the reader is familiar

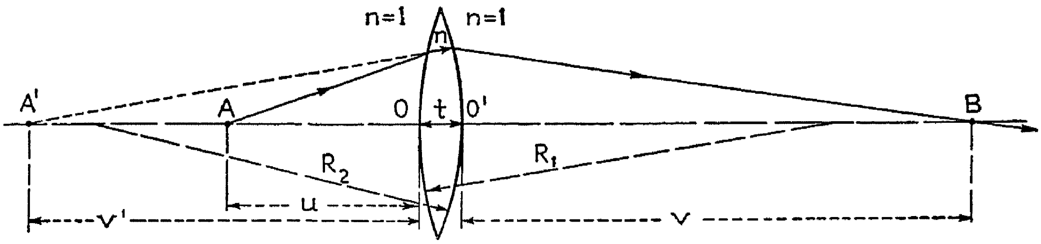

FIG. 171.

with the qualitative behavior of the different types of converging and diverging lenses. In our discussion of the behavior of lenses we shall assume that the apertures are sufficiently small or that aberrations have been corrected, so that Eqs. (13) and (15) may be used to describe the refraction at the lens surfaces.

For the sake of simplicity, we shall examine in this section the so-called *thin* lens, *i.e.*, one for which the thickness t between the vertices of the lens surfaces is small compared to the object and image distances entering into the discussion. In Fig. 171 is shown such a thin double convex lens with surfaces of radii of curvature R_1 and R_2 as shown.

An object A will be imaged by the left-hand lens surface at the point A', and, according to Eq. (13), we have

$$\frac{1}{u} + \frac{n}{v'} = \frac{n-1}{R_1} \tag{16}$$

where v' is the image distance in the image space of index n for this surface and is negative as shown in the figure. This image

A' now serves as the object for the right-hand lens surface in an object space of index n. The object distance $A'O' = u'$ is positive according to our conventions, and for a *thin* lens we may place it equal to $-v'$, neglecting t compared to v'. Thus, applying Eq. (13) to the second surface, we find

$$-\frac{n}{v'} + \frac{1}{v} = \frac{1 - n}{R_2} \tag{17}$$

Since R_2 is negative and since $n > 1$, the right-hand side of Eq. (17) is positive. If we now add Eqs. (16) and (17), we find

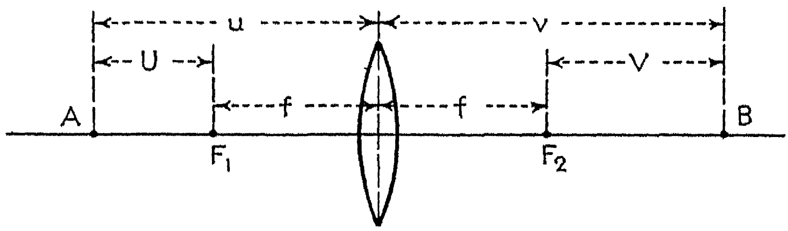

FIG. 172.

the usual equation for the object-image distances for a thin lens. This is

$$\frac{1}{u} + \frac{1}{v} = (n - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) = \frac{1}{f} \tag{18}$$

where we have set

$$\frac{1}{f} = (n - 1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) \tag{19}$$

and u and v are measured from either vertex or from the center of the lens. f is known as the *focal length* of the lens, and it depends only on the material and dimensions of the lens. Positive values of f correspond to converging lenses and negative values to diverging lenses. If an object is placed at a distance f to the left of the lens [$u = f$ in Eq. (18)], the emerging rays will form a parallel beam, and we say that an image is formed at infinity. This point is called a principal *focus* or *focal point* F_1 of the lens. Similarly, parallel rays incident on the lens from the left (corresponding to an object at infinity) are focused at the second principal focus of the lens, F_2, a point which lies at a distance f to the right of the lens. These principal foci are shown in Fig. 172. Any corresponding object and image points, such as A and B, are called *conjugate points*.

There is an important and interesting way of rewriting Eq. (18) (the so-called Newtonian form) which, in many respects, is more useful than the equation we have derived. This form of the equation is obtained by using the focal points of the lens as origins, F_1 for object space, and F_2 for image space, instead of the vertices of the refracting surfaces. If we denote by U the object distance measured from the first focal point and by V the image distance measured from the second focal point, we have from Fig. 172

$$\left.\begin{aligned} U &= u - f \\ V &= v - f \end{aligned}\right\} \tag{20}$$

Substituting in Eq. (18), we find

$$\frac{1}{U+f} + \frac{1}{V+f} = \frac{1}{f}$$

or simplified

$$UV = f^2 \tag{21}$$

This is the Newtonian form of the lens equation.

The lateral magnification of a thin lens may be readily computed with the help of the results of the preceding section. If the linear dimension of an object at A in Fig. 171 is y, the corresponding linear dimension of the image formed by the first surface at A' is, according to Eq. (15),

$$y' = y\left(-\frac{v'}{nu}\right) = m_1 y$$

Similarly, the corresponding linear dimension y'' of the image of A' formed by the second surface at B is

$$y'' = y'\left(-\frac{nv}{A'O'}\right) = y'\left(\frac{nv}{v'}\right) = m_2 y'$$

Thus the magnification of the lens is

$$m = \frac{y''}{y} = -\frac{v}{u} \tag{22}$$

In any case we have the relation

$$m = m_1 m_2 \tag{23}$$

If we now express the magnification in terms of the focal distances U and V, we have from Eqs. (22) and (20)

$$m = -\frac{v}{u} = -\frac{V + f}{U + f}$$

or the more convenient form, using Eq. (21),

$$m = -\frac{f}{U} = -\frac{V}{f} \tag{24}$$

It is customary in optometry to speak of the "power" of a lens instead of its focal length. The power of a lens is defined as the reciprocal of the focal length, and the conventional unit is called a *diopter*, which is 1 meter^{-1}. Since the lens power $1/f$, according to Eq. (19), is the sum of two terms, each corresponding to one of the lens surfaces, one speaks also of the "power" of a lens surface. The sum of the surface powers is then the power of the lens in accordance with the above-mentioned equation.

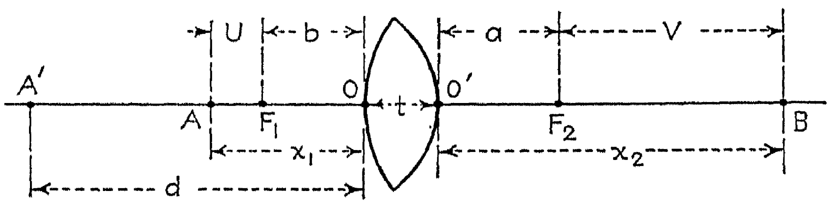

FIG. 173.

68. The Thick Lens in Air.—We shall now drop the restriction that the axial thickness of the lens be small compared to object and image distances from the lens vertices, and we shall analyze the behavior of a thick lens in air for paraxial rays. For convenience we redraw Fig. 171, relabeling the distances as shown in Fig. 173. The object distance from the vertex O of the left-hand lens surface (radius of curvature R_1) is now denoted by x_1, and the image distance from the vertex O' of the right-hand lens surface (radius of curvature R_2) is now x_2. The relation between the object distance x_1 and the image distance $d = A'O$ for the first lens surface is, according to Eq. (13),

$$\frac{1}{x_1} + \frac{n}{d} = \frac{n-1}{R_1} = \frac{1}{f_1} \tag{25}$$

where f_1 is used as an abbreviation for $R_1/(n - 1)$. Using A' as an object for the second surface, the relation between d and x_2 is, using Eq. (13) again and remembering that d is negative for the case of Fig. 173,

$$\frac{n}{t-d}+\frac{1}{x_2}=\frac{1-n}{R_2}=\frac{1}{f_2} \tag{26}$$

where f_2 is defined as $R_2/(1-n)$. For the double convex lens of Fig. 173 both f_1 and f_2 are positive, but our results will hold equally well for other types of lenses. Equations (25) and (26) may be written in the forms

$$d=\frac{nf_1x_1}{x_1-f_1} \tag{25a}$$

$$t-d=\frac{nf_2x_2}{x_2-f_2} \tag{26a}$$

Eliminating d between these equations, one finds readily

$$x_1x_2-ax_1-bx_2-c=0 \tag{27}$$

where we have placed

$$a=\frac{f_2\left(f_1-\frac{t}{n}\right)}{f_1+f_2-\frac{t}{n}};\qquad b=\frac{f_1\left(f_2-\frac{t}{n}\right)}{f_1+f_2-\frac{t}{n}};\qquad c=\frac{f_1f_2\frac{t}{n}}{f_1+f_2-\frac{t}{n}} \tag{28}$$

In these relations t is essentially a positive quantity. Writing Eq. (27) in the form

$$(x_1-b)(x_2-a)=ab+c \tag{29}$$

we see that the first principal focus of the lens, F_1, lies at a distance b to the left of the vertex O as shown in Fig. 173. For the object A placed at F_1, $x_1=b$ and $x_2=\infty$ according to Eq. (29). Similarly, the second principal focus F_2 lies at a distance a to the right of the vertex O', since an object at infinity ($x_1=\infty$), corresponding to parallel rays incident on the left-hand lens surface, gives rise to an image at $F_2(x_2=a)$. We have thus found the positions of the focal points of the thick lens. The planes perpendicular to the lens axis through these focal points F_1 and F_2 are called the *focal planes* of the lens.

Now let us measure object and image distances from the focal planes at F_1 and F_2, respectively, and denote them by U and V, just as we did in the case of the thin lens. We have evidently (see Fig. 173)

$$\left.\begin{aligned} U &= x_1 - b = x_1 - \frac{f_1\left(f_2 - \dfrac{t}{n}\right)}{f_1 + f_2 - \dfrac{t}{n}} \\ V &= x_2 - a = x_2 - \frac{f_2\left(f_1 - \dfrac{t}{n}\right)}{f_1 + f_2 - \dfrac{t}{n}} \end{aligned}\right\} \tag{30}$$

and Eq. (29) takes the familiar Newtonian form

$$UV = ab + c = \frac{f_1^2 f_2^2}{\left(f_1 + f_2 - \dfrac{t}{n}\right)^2} = f^2 \tag{31}$$

utilizing Eq. (28) and defining f^2 by this equation. The positive square root of this expression for f^2 is called the *focal length* of the lens, and we have

$$f = \frac{f_1 f_2}{f_1 + f_2 - \dfrac{t}{n}} \tag{32}$$

The Newtonian form of the lens equation which we have derived for the thick lens suggests strongly that it should be possible to obtain an equation of the form of Eq. (18) for the thin lens. This is possible by proper choice of origins, and these new reference points from which one may measure object and image distances are called the *principal points* of the lens and are denoted by H_1 and H_2, respectively. We determine the positions of these principal points as follows: Let U_1 and V_1 be the coordinates of the principal points H_1 and H_2 in object and image space, respectively, referred to the focal points as origins. If, then, u and v are object and image distances measured from the principal points, we have evidently

$$\left.\begin{aligned} u &= U - U_1 \\ v &= V - V_1 \end{aligned}\right\} \tag{33}$$

where U and V are the object and image distances from the focal points. If now the thin lens equation

$$\frac{1}{u} + \frac{1}{v} = \frac{1}{f}$$

is to be valid, we must have

$$\frac{1}{U - U_1} + \frac{1}{V - V_1} = \frac{1}{f}$$

or, rewritten,

$$fV - fV_1 + fU - fU_1 = f^2 - UV_1 - VU_1 + U_1V_1$$

where we have used the fact that $UV = f^2$. This equation is evidently satisfied if we place

$$\left.\begin{aligned} U_1 &= -f \\ V_1 &= -f \end{aligned}\right\} \tag{34}$$

so that the principal points are also conjugate points of the lens. We thus have the important result that the thin lens equation holds equally well for a thick lens, provided the principal points H_1 and H_2 are used as origins for object and image distances, respectively. Clearly the separation between either focal point and the corresponding principal point is the focal length of the lens.

Let us compute the positions of the principal points (or planes) relative to the vertices of the lens surfaces which we used originally as reference points. x_1 is the object distance from the left-hand vertex O, and u is the object distance from the first principal point H_1. The relation between u and x_1 is given by

$$u = U - U_1 = x_1 - b + f = x_1 + \frac{f_1\dfrac{t}{n}}{f_1 + f_2 - \dfrac{t}{n}} \tag{35}$$

using Eqs. (30), (32), and (34).

Similarly, the image distance v as measured from the second principal point H_2 is related to the image distance x_2 measured from the right-hand vertex O' by

$$v = V - V_1 = x_2 - a + f = x_2 + \frac{f_2\dfrac{t}{n}}{f_1 + f_2 - \dfrac{t}{n}} \tag{36}$$

Note that for a thin lens $u = x_1$ and $v = x_2$, so that the principal points coincide and lie at the center of the lens. Equations (35) and (36) locate the principal points of a thick lens with

respect to its vertices. Thus H_1 lies at a distance

$$h_1 = -\frac{f_1\frac{t}{n}}{f_1 + f_2 - \frac{t}{n}} \tag{37}$$

from the left-hand vertex in object space (in our case to the right of O), and H_2 lies at a distance

$$h_2 = -\frac{f_2\frac{t}{n}}{f_1 + f_2 - \frac{t}{n}} \tag{38}$$

from the right-hand vertex in image space (in our case to the left of O').

For the sake of completeness, we redraw the lens of Fig. 173, showing the principal points, focal points, focal length, and object and image distances (Fig. 174).

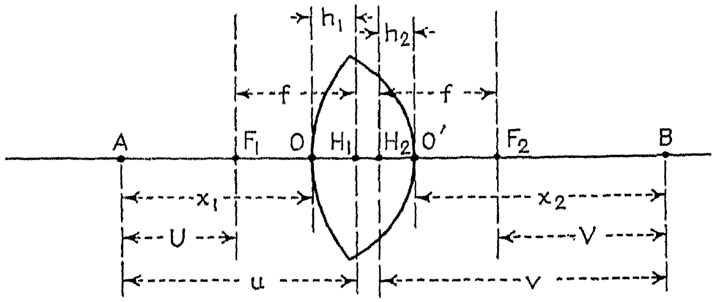

FIG. 174.

Not only the object-image distance equations but also the magnification formulas are identical in form for thick and thin lenses. The magnification of the lens is the product of the magnifications of the two surfaces, and this is, using Eq. (15) (see Fig. 173),

$$m = \left(-\frac{d}{nx_1}\right)\left(-\frac{nx_2}{t-d}\right)$$

From Eqs. (25*a*) and (26*a*), we have

$$\frac{d}{t-d} = \frac{x_2 - f_2}{x_1 - f_1} \cdot \frac{f_1 x_1}{f_2 x_2}$$

so that

$$m = \frac{f_1}{f_2} \cdot \frac{x_2 - f_2}{x_1 - f_1} = \frac{f_1}{f_2} \frac{V + a - f_2}{U + b - f_1}$$

the last equality following from Eq. (30).

Using the definitions of a and b and of the focal length f [Eqs. (28) and (32)], one finds readily that

$$a - f_2 = -\frac{f_2}{f_1}f \qquad \text{and} \qquad b - f_1 = -\frac{f_1}{f_2}f$$

so that

$$m = \frac{f_1}{f_2} \cdot \frac{V - \dfrac{f_2}{f_1}f}{U - \dfrac{f_1}{f_2}f}$$

and, since $UV = f^2$ (Eq. 31), this becomes

$$m = -\frac{f}{U} = -\frac{V}{f} \tag{39}$$

which is identical with Eq. (24). It immediately follows that Eq. (22) also holds for the thick lens, remembering that u and v are measured from the principal points. Equation (39) for the magnification holds for all object-image pairs, U and V being object and image distances as measured from the focal points F_1 and F_2. For the special case for which object and image lie at the principal points $U = V = -f$, we have unit magnification and a virtual image.

It is a property of the principal points that an incident ray through one of them emerges through the other parallel to the incident ray. The proof of this is left to a problem. Finally, it should be pointed out that all our results are valid for any lens or combinations of lenses, provided the object and image space have the same index of refraction. If this is not so, then only the form of the magnification formulas becomes modified, and the conjugate points for unit magnification (the so-called *nodal points*) no longer coincide with the principal points.

69. Lens Aberrations.—Thus far our considerations have been restricted largely to what may be termed the "first-order" theory of the behavior of both thin and thick lenses. There have been two fundamental assumptions in the theory: (1) The objects and images have been treated as if they were situated on the axis of the lens, so that at best the theory gives a good

approximation for objects and images, all the points of which lie very close to the axis; and (2) the rays from a point object on the axis have been considered as a very narrow pencil, so that the angular opening of the cone of rays is small enough to allow us to replace the sine of the angle by the angle. Thus we speak of "paraxial" rays.

In general, a lens must image points not on its axis, and the cone of rays from any point of the object will be of finite angular opening. Furthermore, even if individual object points were sharply focused at corresponding image points, we might still obtain distortion of the image in the sense that the geometrical shape of the image would not coincide with that of the object. Finally, we have dealt with the index of refraction as a constant. While this is true for light of a given frequency (monochromatic light), white light is a mixture of waves of many wave lengths and the index of refraction of glass, for example, varies with wave length (and frequency). Thus, even in the first-order theory, one encounters so-called *chromatic aberration*, variation of focal length with color of the light. The various defects of the image formed by a perfectly spherical lens are termed *aberrations*, and we shall briefly examine these various types of aberrations. A quantitative study of these aberrations and methods of minimizing them are far beyond the scope of this book. For the sake of simplicity we shall consider the aberrations for the case of a thin lens in air, although the same general considerations hold for thick lenses and combinations of lenses.

We classify the various aberrations as follows:

1. If a narrow pencil of rays (parallel or stigmatic) lies along the lens axis, we have the conditions of our first-order theory satisfied and have no aberration.

2. If a narrow pencil of rays passes *obliquely* through the lens, as does a portion of the cone of rays from a point object not located on the lens axis, this pencil becomes *astigmatic* after it passes through the lens. This astigmatic pencil does not focus at a point, but all the rays pass through two mutually orthogonal lines which are displaced from each other (focal lines). This aberration is called *astigmatism*.

3. If a wide pencil of rays has its axis coincident with that of the lens or if we have a broad beam of parallel rays which are parallel to the lens axis, the rays of the pencil emerging from the lens will not pass through a single point. This aberration is called *axial spherical aberration*.

4. If a wide stigmatic pencil of rays (or broad parallel beam) passes obliquely through a lens, it becomes greatly confused upon emerging from the lens, and this type of aberration is denoted as *comatic aberration*, or simply as *coma*.

5. For objects of finite size, *e.g.*, an object in the form of a straight line perpendicular to the lens axis, there will result an image which is curved, and this aberration is known as *curvature of the field*. It is intimately connected with astigmatism, being essentially due to the astigmatic nature of the narrow pencils of rays from the various points of the object.

6. *Image distortion* is an aberration arising from a variation of magnification with distance of the various object points from the lens axis.

7. *Chromatic aberration*, as we have already mentioned, is due to the variation of index of refraction with the frequency of the light and results in the formation of a series of images at different distances and of different magnifications, one for each color present in the incident light.

We thus have classified six types of aberration of a simple thin lens. It should be kept in mind, however, that the monochromatic aberrations are classified into the types 2–6 for the sake of convenience, but they are not

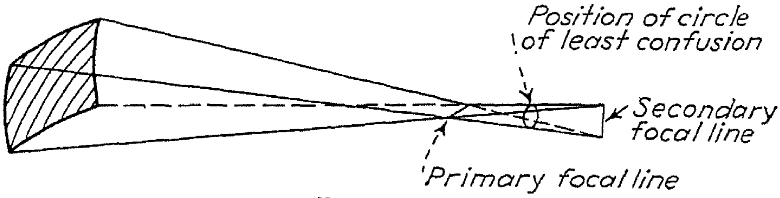

FIG. 175.

independent, and in general it is only approximately correct to consider any one type as being present to the exclusion of all the others. We shall now describe the above aberrations in more detail.

Astigmatism and Curvature of the Field.—Let us first remind ourselves as to the properties of an astigmatic pencil of rays. These rays are the normals to a small portion of a surface of constant phase which has two different radii of curvature in two mutually orthogonal directions. The pencil of rays is indicated in Fig. 175. All the rays of the pencil pass through two

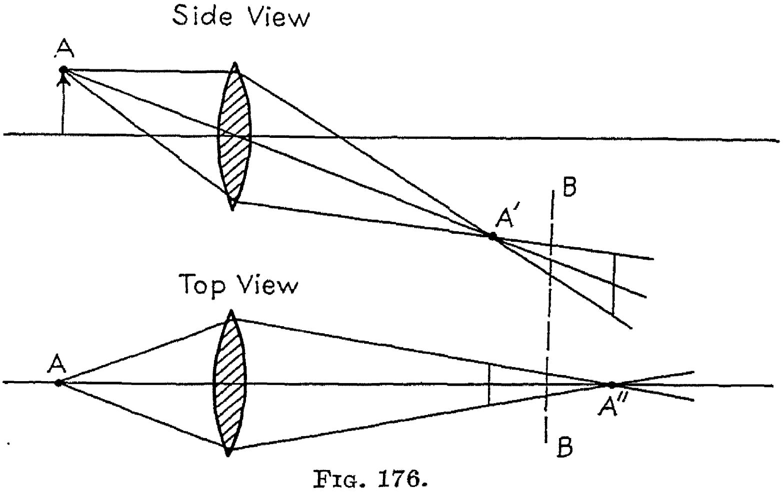

FIG. 176.

focal lines, as shown in the figure. The cross section of the pencil is in general elliptical for an elliptical boundary on the wave surface, the ellipse degenerating into straight lines at the positions of the focal lines. Between these two positions there is a position where the cross section of the beam is circular, the major and minor axes of the ellipse becoming equal, and this circle is known as the *circle of least confusion*, giving the best image formed by an astigmatic pencil. As we have stated, a narrow pencil of rays passing

obliquely through a thin lens gives rise to an astigmatic transmitted pencil. In Fig. 176 are shown two views (top and side) of such a pencil of rays (exaggerated for the sake of clarity) from the tip of an arrow used as an object for a thin lens. A' and A'' are the locations of the focal lines of the pencil, and BB is the location of the circle of least confusion. A' and A'' are often called the primary and secondary images of A.

If one considers all the points of the arrow, instead of only the tip, the locus of all the primary images will form a curve, as will the locus of the secondary images. These two curves will be tangent to each other at the axis, and somewhere between them lies a surface containing the circles of least confusion for all the points of the object. This is the surface of *best focus*, and in general it is not a plane, giving rise to the curvature of the best possible image of the object. The shape of the image surfaces depends on the shape of the lens and on the positions of stops or apertures on the lens axis. It is not possible to eliminate both curvature of the field and astigmatism

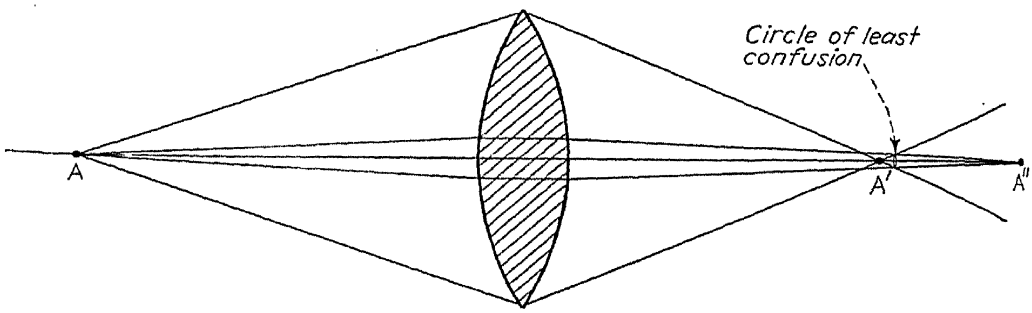

FIG. 177.

with a single lens, but either may be readily corrected. A narrow aperture lens corrected for astigmatism is called *anastigmatic*.

Spherical Aberration.—In Fig. 177 is shown the effect of spherical aberration. A finite cone of rays from a point A on the lens axis is not imaged at a single point. All the rays making a given angle with the axis are imaged at the same point, the widest angle rays at A' and the small angle or paraxial rays at A''. The cross section of the beam is everywhere circular, and the circle of least radius is the circle of least confusion, the best image of A. Spherical aberration may be reduced by "stopping down" the lens (using only the central portion), but only at the expense of the amount of light transmitted by the lens. A lens of given focal length may be designed for minimum spherical aberration by choosing appropriate radii of curvature for the two surfaces. In general, this type of aberration is cut down by dividing the deviation of the rays equally between the two surfaces.

Coma.—The image of a point object not on the lens axis, when the lens intercepts a cone of rays of large angle, is very confused and is called comatic. Let us assume that astigmatism may be neglected, so that there is a well-defined plane of best focus. The finite cone of rays from the point object may be considered as a sum of very narrow pencils with a common focal point. As we have seen, each of these narrow pencils gives a best image in the form of a circle of least confusion. In general, the centers and radii of

these circles of least confusion will be different for the different narrow pencils making up the cone of rays, so that the final image may be thought of as a superposition of circles of different radii and centers. The resultant figure is shaped somewhat like a comet (hence the name coma) and is illustrated in Fig. 178. Coma may be eliminated completely for a single thin lens for *one pair* of object-image points by proper choice of the radii of curvature of the lens surfaces. Such a lens will then necessarily show axial

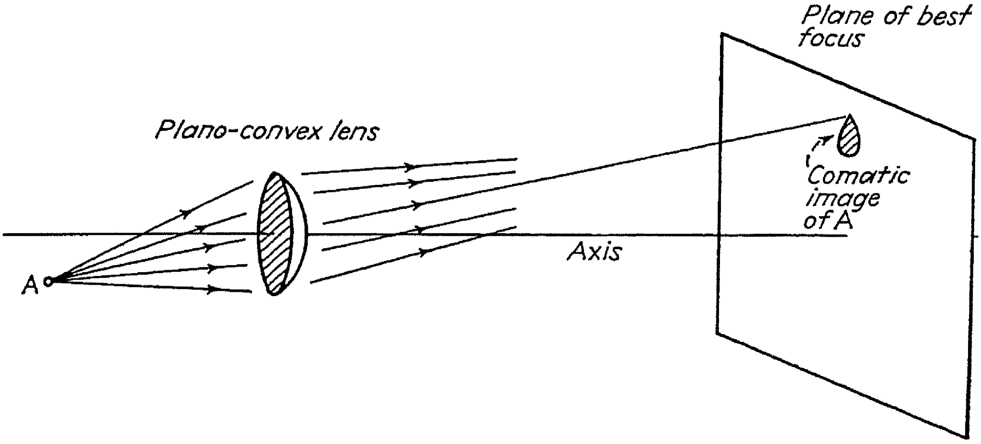

FIG. 178.

spherical aberration, since the condition for no coma is not the same as for minimum spherical aberration. A very wide aperture lens may be corrected for axial spherical aberration and for coma for slightly oblique rays. It is then said to be *aplanatic*, and the single pair of object and image points for which this correction is valid are called the *aplanatic points* of the lens.

Image distortion, caused by the variation of magnification with lateral distance of an object point from the lens axis, may be of two types, depending on whether the magnification increases or decreases with distance from

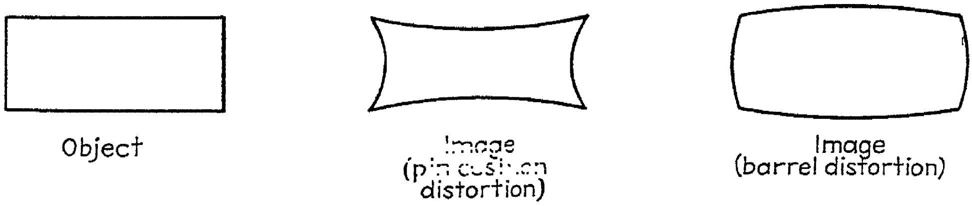

FIG. 179.

the axis. In the former case the distortion is known as "pincushion" distortion and in the latter case as "barrel" distortion. A rectangular object will give rise to the two images shown in Fig. 179 for these two types of distortion.

Chromatic aberration and the methods of minimizing it will be discussed in the chapter on dispersion of light.

It is clear that it is impossible to eliminate or even to minimize simultaneously all the aberrations discussed above for a single lens, but it is possi-

ble, by combining a number of lenses to form a *compound lens*, to balance the aberrations of one part of the system against the other. The larger the number of elements of a compound lens, the higher the degree of correction which may be obtained. In practice, attention is directed toward correcting those aberrations which would be most disturbing for the particular purpose to which the lens is put. When we discuss chromatic aberration, we shall illustrate the manner in which a compound lens may be designed to compensate for chromatic aberration.

70. The Eye.—The essential parts of the human eye, treated as an optical instrument, are shown in Fig. 180. The eye is roughly spherical, the front surface of the eyeball being somewhat more sharply curved and covered with a tough transparent membrane C, the *cornea*. Between the cornea and the *lens* L is a watery liquid known as the *aqueous humor*. The remainder of the eyeball is filled with a transparent viscous liquid, known as the *vitreous humor*. The front surface of the cornea and the lens surfaces are nearly spherical and constitute essentially a compound lens which projects images of external objects upon a sensitive membrane, the retina R, in which terminate the ends of a great many nerve fibers entering the eye at O, the nerve bundle being known as the optic nerve. The *pupil* P is an aperture in a muscular membrane called the *iris* and serves to cut down or increase the amount of light entering the eye.

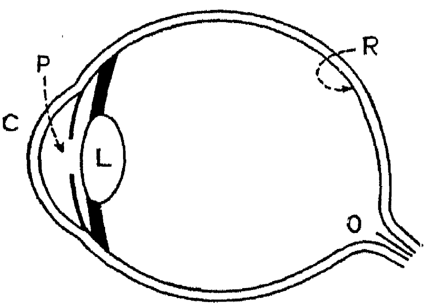

FIG. 180.

The lens of the eye may be compressed into a more nearly spherical shape by the action of a muscle attached to it and thus may have its focal length varied. The process of focusing the eye on objects at varying distance from it is called *accommodation*, and the range over which distinct vision is possible determines the so-called *near* and *far points* of the eye. For the normal eye the far point may be taken as infinity. The position of the near point depends on how much the curvature of the lens may be increased by accommodation, and the power of accommodation diminishes with increasing age. For the normal eye we shall take the near point as about 15 cm., and the distance of *most distinct vision* as 25 cm. (or 10 in.).

71. The Simple and the Compound Microscope.—When looked at directly with the eye, an object seems large when the retinal

image is large, and the angle subtended at the eye by the object (the so-called *visual angle*) is a measure of the apparent size of the object. In Fig. 181 the angle θ is the visual angle. In order to examine an object in detail, it is brought as near to the eye as possible to obtain a large visual angle. Since the eye

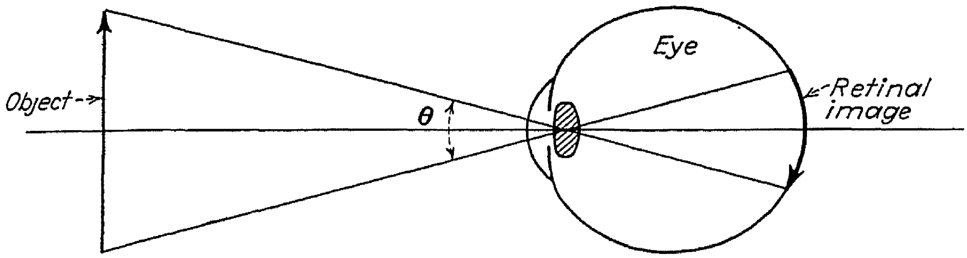

FIG. 181.

cannot focus sharply on objects closer than the near point, the maximum visual angle obtainable by the unaided eye is limited by the power of accommodation. By placing a converging lens in front of the eye, one can in effect increase its effective accommodation. The eye then looks at an enlarged virtual image as indicated in Fig. 182. Such a device is known as a magnifying

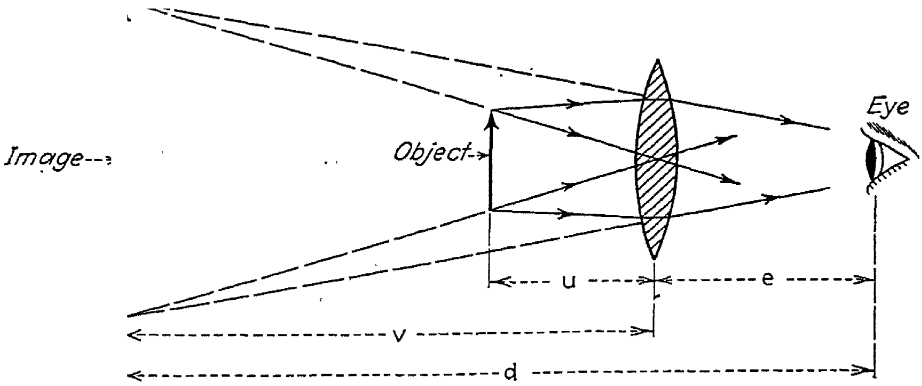

FIG. 182.

glass or simple microscope. *The magnifying power M of a magnifying glass is defined as the ratio of the angle subtended at the eye by the image to that subtended by the object when the latter is placed at the distance of most distinct vision d_0, which we shall take as 25 cm.*

If θ is the angle subtended at the eye by the image of Fig. 182 and y_i its length, we have

$$\frac{y_i}{2} = d \tan\left(\frac{\theta}{2}\right) \tag{40}$$

If the object has a length y_0 and is located at the distance of most distinct vision d_0, it subtends an angle θ_0 at the eye, where

$$\frac{y_0}{2} = d_0 \tan\left(\frac{\theta_0}{2}\right) \tag{41}$$

For small angles we may replace the tangents of the angles by the angles and have for the magnifying power

$$M = \frac{\theta}{\theta_0} = \frac{y_i}{y_0} \cdot \frac{d_0}{d} \tag{42}$$

We have further from Fig. 182

$$\frac{y_i}{y_0} = \frac{d - e}{u} \tag{43}$$

so that Eq. (42) becomes

$$M = \left(1 - \frac{e}{d}\right)\frac{d_0}{u} \tag{44}$$

In general, the eye is placed close enough to the magnifier that the distance e in Eq. (44) may be neglected compared with d, and hence we have simply

$$M = \frac{d_0}{u} \tag{45}$$

If the image is formed at the distance of most distinct vision, we have $v = -d_0$, and from the lens equation we have

$$\frac{1}{u} - \frac{1}{d_0} = \frac{1}{f}$$

so that the magnifying power becomes for this case

$$M = \frac{d_0}{u} = \frac{d_0}{f} + 1 \tag{46}$$

If the image is formed at infinity, we have $\frac{1}{u} = \frac{1}{f}$, so that for this case

$$M = \frac{d_0}{f} \tag{47}$$

Thus the magnifying power is somewhat greater when the image is formed at the point of most distinct vision.

According to the above results, a high-power magnifying glass must have a very short focal length, and it must be held

close to the object under examination and near to the eye. The difficulty of accurately grinding a lens of small focal length and freeing it as far as possible from aberrations, in addition to the inconveniences of operation just mentioned, place a practical upper limit of about 20× on the magnifying power of a simple microscope. For higher powers one utilizes a *compound microscope*, consisting essentially of a lens, called the *objective*, which produces a real enlarged image of the object, and a magnifying glass or ocular for viewing this image. The arrangement is shown in Fig. 183 in which both objective and ocular are treated as simple lenses, although in an actual microscope both would be highly corrected compound lenses. The object O is placed

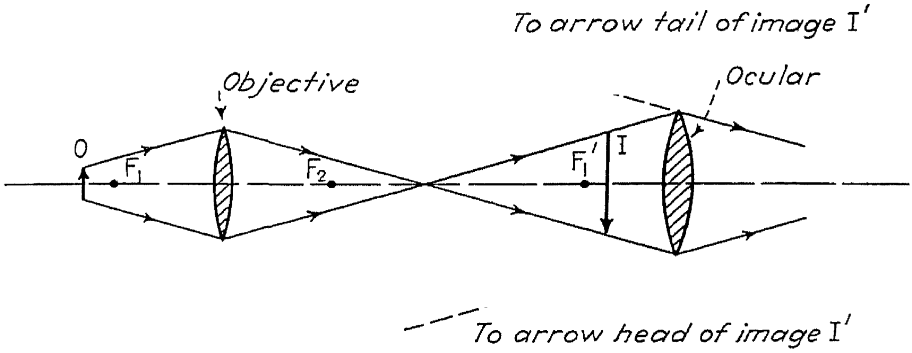

FIG. 183.

just outside the first focal point F_1 of the objective, yielding an enlarged real image I. This image lies just inside the first focal point F_1' of the ocular, giving rise to a virtual enlarged image I' of I (not shown in the figure). The position of I' may be taken at the point of most distinct vision of the eye.

The magnifying power of the compound microscope is the product of the magnifying powers of the objective and ocular. Since the objective forms a real image which is examined by the ocular, its magnifying power is just the linear magnification m_1, which, according to Eq. (39), may be written as

$$m_1 = -\frac{\quad}{f_1} \tag{48}$$

where V is the focal distance of the image I (the distance from F_2 to I) and f_1 the focal length of the objective. The magnifying power of the ocular is, according to Eq. (46),

$$M_1 = \frac{d_0}{f_2} + 1 \tag{49}$$

where f_2 is the focal length of the ocular. Thus the magnifying power of a compound microscope is

$$M = m_1 M_1 = \left(-\frac{V}{f_1}\right)\left(\frac{d_0}{f_2} + 1\right) \tag{50}$$

or

$$M = m_1 M_1' = -\left(\frac{V}{f_1}\right)\left(\frac{d_0}{f_2}\right) \tag{50a}$$

the latter equation being valid when I' is formed at infinity. It has become standard practice to make $V = 18$ cm., so that, using the value of 25 cm. for d_0, we may write in place of Eq. (50a)

$$M = -\left(\frac{18}{f_1}\right)\left(\frac{25}{f_2}\right) \tag{51}$$

where f_1 and f_2 are expressed in centimeters.

72. Oculars.—In actual practice the magnifying glass shown in the compound microscope of Fig. 183 is replaced by a com-

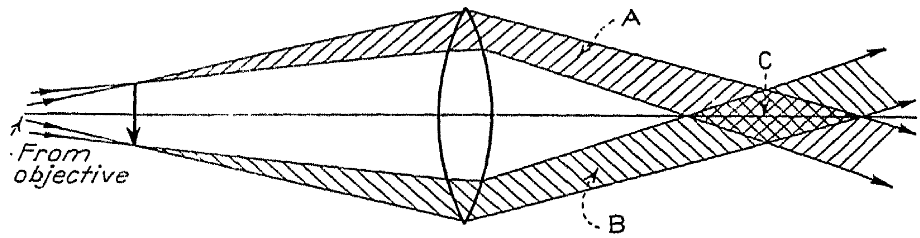

FIG. 184.

pound lens. When a compound lens is used to examine the image formed by a system of lenses (or mirrors), it is called an *ocular* or *eyepiece.* In addition to the primary purpose of magnification, a well designed ocular increases the *field of view* and, of course, lends itself to a better reduction of aberrations than a simple lens. The fundamental difference between a real object and an image used as an object lies in the fact that the rays from the former diverge in all directions, whereas in the latter the pencil of rays from any point forms a cone of limited opening. This is illustrated in Fig. 184 in which are indicated the cone of rays coming from the head and tail of an image which is being examined by a simple magnifier. If the eye is placed at A, it is clear that none of the rays from the arrowhead reach it, so that the arrowhead lies outside the field of view. Similarly at B the tail of the arrow is not visible. By placing the eye at C, one

can see the whole arrow, but in most instruments it is desirable to have the eye close to the eyepiece.

The method employed to increase the field of view for an eye placed near the eyepiece consists essentially of the use of two lenses in the ocular. The function of the first lens, called the *field lens*, is to alter the direction of the rays from the various points of the image under examination, so that pencils of rays from each of these points pass through the central portion of the second or *eye lens* and hence can all enter the pupil of an eye placed near this eye lens. Two common types of oculars are shown in Figs. 185 and 186. In the Ramsden ocular the lenses are of equal focal length and the image is formed in front of the

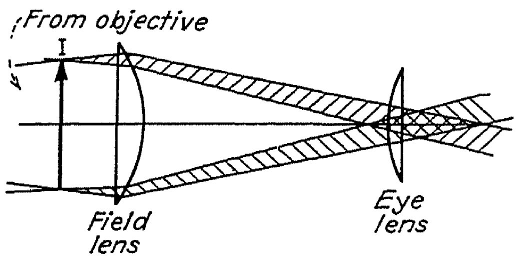

FIG. 185.—Ramsden ocular.

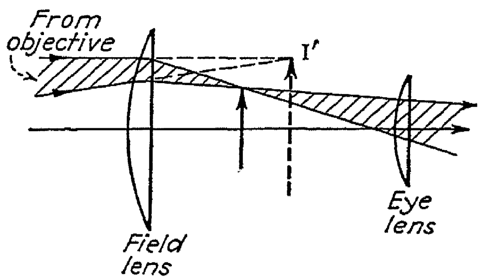

FIG. 186.—Huygens ocular.

field lens as shown. If cross hairs or a scale are to be mounted in the ocular, so that measurements may be made, they should be placed in the plane of the image I. In the Huygens ocular the field lens intercepts the rays from the objective before the image is formed, and thus we have a virtual object I' for this field lens, giving rise to the image I, which is then examined with the help of the eye lens. Clearly the Ramsden ocular may be used to examine a real object but the Huygens ocular cannot be so employed.

73. Telescopes.—The simple astronomical telescope consists essentially of an objective converging lens of long focal length and an ocular for examining the image of a distant object formed by the objective. A schematic sketch of the optical system is given in Fig. 187, in which no attempt has been made to show the paths of the rays. Since the object is practically at infinity, the image I is formed at the second focal point of the objective, and, if the image formed by the ocular of I is also at infinity, I lies at the first focal point of the ocular. The angles θ and θ' are the angles subtended by the object at the objective and by

the image at the ocular. *The magnifying power M of a telescope is defined as the ratio of these angles*, and, since in the telescope shown we have an inverted image, we write

$$M = -\frac{\theta'}{\theta}$$

and, because in general the angles are small, this can be written in the form

$$M = -\frac{f}{f'} \tag{52}$$

as is evident from Fig. 187, where f is the focal length of the objective and f' that of the ocular.

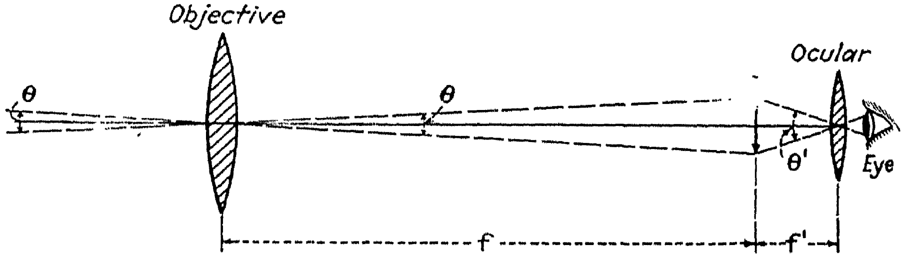

FIG. 187.

The inverted image formed by the simple telescope described above is not disadvantageous for astronomical work, but for terrestrial purposes it is desirable to have an upright image. This can be accomplished by the addition of an erecting lens to the astronomical telescope as shown in Fig. 188 to form a *terrestrial*

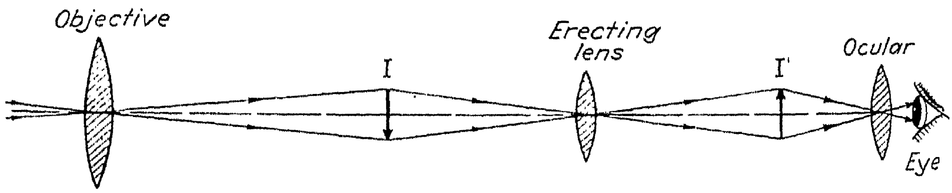

FIG. 188.

telescope or *spyglass*. The disadvantage of this arrangement is that the length of the instrument becomes unwieldy. Since the shortest object-image distance for a thin convex lens which forms a real image is four times its focal length, the minimum length of the telescope tube becomes the sum of the focal lengths of objective and ocular and four times the focal length of the erecting lens.

This disadvantage may be overcome by using a diverging instead of a converging lens for the eyepiece, and a telescope so constructed is called a *Galilean telescope*. The ordinary opera glass is such a telescope. Figure 189 shows a diagram of the elements. It is readily seen that the magnifying power M is

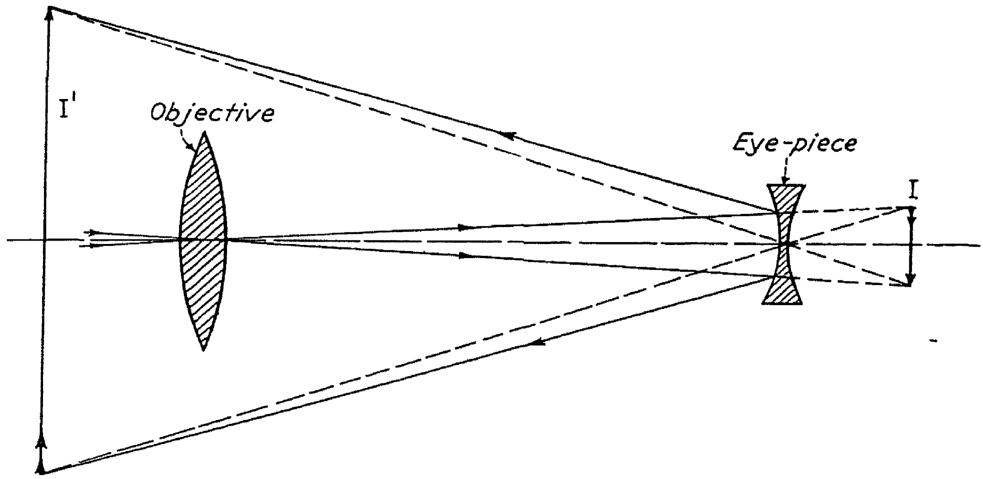

FIG. 189.

given by Eq. (52) $\left(M = -\frac{f}{f'}\right)$, and, since f' is negative, this indicates the upright image. The length of the telescope is the difference of the focal lengths of the two lenses instead of the sum, as in the case of the astronomical telescope, thus yielding a compact instrument.

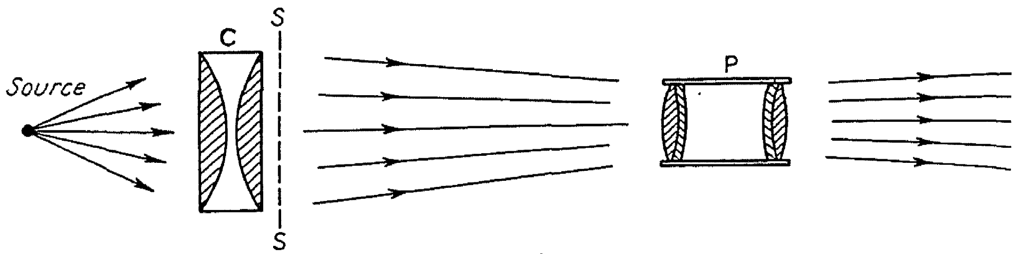

FIG. 190.

74. The Projection Lantern.—Figure 190 shows the essential parts of the optical system of a projection lantern. The condensing lens C, usually consisting of a pair of plano-convex lenses as shown, serves to deviate the rays from the source through the lantern slide SS so that they are all intercepted by the projection lens P. Its function is thus very similar to that of the field lens in an ocular. The projection lens usually

consists of two separate lenses as shown, each of which is a compound lens with two kinds of glass (to correct for chromatic aberration as we shall see later) and must be corrected for curvature of the field and for image distortion. Since the final image is produced by the projection lens only, it is not necessary to correct for aberrations in the condensing lens.

The simple optical instruments which we have discussed are all of the image-forming type, and we postpone discussion of the so-called analyzing instruments, *i.e.*, those employed to determine the spectral composition or intensity of a beam of light, until we have occasion to study the results obtained by their use.

Problems

1. What is the apparent depth of a swimming pool in which there is water of depth 6 ft.,

a. When viewed from above at normal incidence?

b. When viewed at an angle of 60° with the surface?

The index of refraction of water is 1.33.

2. A layer of ether (n = 1.36) 2 cm. deep floats on water (n = 1.33), 4 cm. deep. What is the apparent distance from the ether surface to the bottom of the water when viewed at normal incidence?

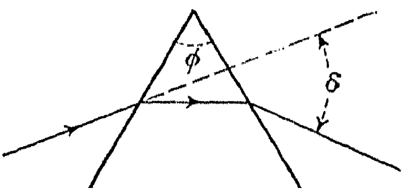

FIG. 191.

3. To measure the refractive index of a crystal in the form of a flat plate 12.62 mm. thick, a microscope is focused on a scratch on the top surface and then lowered a distance of 5.43 mm. to focus on a scratch on the bottom surface. Compute the refractive index of the crystal.

4. Derive the law of reflection using Fermat's principle.

5. Light is deviated by a glass prism of index n as shown in Fig. 191. The ray in the prism is parallel to the base. Show that the index of refraction is related to the angle of deviation δ and prism angle ϕ by the equation

$$n = \frac{\sin\left(\dfrac{\phi + \delta}{2}\right)}{\sin\,(\phi/2)}$$

for this angle of incidence. The deviation angle δ is a minimum for this angle of incidence and is known as the angle of minimum deviation.

6. What is the relation between the size of an observer and the minimum size of a plane mirror, if the observer is to see his complete image?

7. A concave spherical mirror has a radius of curvature of 50 cm. Find two positions of an object such that the image be four times as large as the object. What is the position of the image in each case? Is it real or virtual?

8. Derive the appropriate form of Eq. (7) of the text for the case of a convex mirror.

9. A convex mirror has a focal length of 10 in. Compute the position of the image of an object 6 in. in front of the mirror. What is the magnification for this case?

10. A glass rod, index of refraction n, has its ends ground spherically, the radii being R_1 and R_2 as shown in Fig. 192. The length of the rod is $l = 2R$, and a small object is placed at A. Show that the image of A lies at a distance $R\left(\frac{4-n}{3n-4}\right)$ from the right-hand vertex O'. (Assume small angles of incidence.)

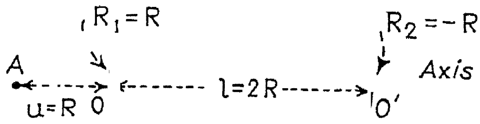

FIG. 192.

11. A narrow pencil of parallel light rays is normally incident on a solid glass sphere of radius R and refractive index n. How far from the center of the sphere are the rays brought to a focus?

12. A narrow pencil of parallel light rays is normally incident on a hollow glass sphere of inner radius 2 in. and outer radius 2.5 in. If the index of refraction of the glass is 1.60, where will these rays be brought to a focus? Make a careful drawing of the pencil of rays.

13. Prove that Eq. (13) of the text is valid for refraction at a surface of negative radius of curvature (concave toward the object).

14. The focal length of a thin lens is determined as 25 cm. by using a lamp as an object, placing the lamp far from the lens, and measuring the image-lens separation. How far away must the lamp be placed to attain a precision of 3 per cent?

15. A thin glass lens of focal length 12 in. and refractive index 1.52 is immersed in water of refractive index 1.33. What is the focal length in water?

16. Prove that, when two thin lenses of focal lengths f_1 and f_2 are placed in contact, they are equivalent to a single lens of focal length f equal to

$$f = \frac{f_1 f_2}{f_1 + f_2}$$

17. Prove that the focal length of a double convex thin lens of diameter d and axial thickness t is given by

$$f = \frac{d^2}{8(n-1)t}$$

where n is the refractive index of the lens material.

18. Derive the appropriate form of Eq. (18) of the text for a double concave lens.

19. A thin double convex glass lens ($n = 1.50$) has a focal length of 3.00 cm. and a diameter of 2.00 cm. An object in the form of an arrow 2.00 cm. long is placed perpendicular to, and bisected by the lens axis at a distance of 5.00 cm. to the left of the lens. Locate the image, and construct a careful diagram of the pencil of rays intercepted by the lens.

What fractional part of this image will be visible to an eye on the lens axis at a point 22.5 cm. to the right of the lens?

20. Make a careful plot of image distance as a function of object distance:

a. For a thin converging lens.

b. For a thin diverging lens.

21. A thin convex and a thin concave lens each of 20 cm. focal length are placed coaxially at a separation of 6 cm. Find the position of the image formed by this lens system of an object at a distance of 30 cm.:

a. Beyond the convex lens.

b. Beyond the concave lens.

22. A thin convex lens of focal length f produces a real image of magnification m. Show that the object distance is

$$u = \frac{m+1}{m} f$$

23. An image of height a is formed on a screen by a thin convex lens. It is found by moving the lens toward the screen that there is a second lens position in which it forms a second sharp image on the screen of height b.

Show that the height of the object is $\sqrt{ab}$.

24. Prove that the distance between an object and its real image formed by a thin convex lens is always greater than four times the focal length of the lens.

25. The radii of curvature of a double convex lens have magnitudes of 20 and 30 cm. The glass of which it is made has a refractive index of 1.52. Find the focal length of the lens.

What is the focal length of a convex meniscus lens with the same radii of curvature and made of the same glass?

26. Each of two similar thin converging lenses has a focal length of 10 in. They are mounted coaxially, and are separated by a distance of 10 in. Find the positions of the images of a small object placed on the axis at a point 20 in. to the left of the first lens.

27. A plano-convex lens of glass of refractive index 1.5 and radius of curvature 24 cm. is 2 cm. thick along the axis. Calculate its focal length, and find the position of the image when the object is 50 cm. from the convex surface:

(*a*) When on the convex side.

(*b*) When on the plane side.

What is the magnification in each of these cases?

28. An object is displaced a small distance du along the axis of a lens. Find an expression for the corresponding displacement dv of the image. Treating du as an object length along the axis and dv as the corresponding image length, the ratio dv/du is often referred to as the *longitudinal magnification*. Prove that it is equal to the square of the ordinary lateral magnification.

29. Prove that the focal length f of a compound lens consisting of two coaxial thin lenses separated by a distance a is given by

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{a}{f_1 f_2}$$

where f_1 and f_2 are the focal lengths of the two lenses, respectively.

30. Prove that, if an incident ray on a thick lens is directed so that it will pass through one of the principal points, the emergent ray has a direction which is parallel to the first and passes through the other principal point.

31. Derive expressions for the positions of the principal points and of the principal foci for two coaxial thin lenses separated by a distance a and of focal lengths f_1 and f_2, respectively.

32. Two thin convex lenses of focal lengths 30 and 10 cm. are mounted coaxially, the 30-cm. focal length lens being on the left. Compute the positions of the principal points and of the principal foci for the following separations; 5 cm., 15 cm., 25 cm., 40 cm., 55 cm., and 70 cm.

Construct a diagram in which these points are located approximately to scale, plotting the six configurations one under the other with the left-hand lens always at the same position.

33. A glass hemisphere of radius 5 cm. and refractive index 1.5 is used as a lens, rays passing through it very nearly coinciding with its axis. Where are the principal points of this lens, and what is its focal length?

34. Sketch the lenses described below and locate the positions of their principal planes. The radius R_1 is that of the left-hand surface, R_2 of the right-hand surface, t the thickness. Take $n = 1.5$.

	Lens a	Lens b	Lens c	Lens d	Lens e
R_1, cm..................	+10	−10	∞	+ 5	−10
R_2, cm..................	−10	+10	+10	+10	− 5
t, cm....................	2	2	1.5	1.5	1.5

35. An image is formed in a space of refractive index n_2 of an object in a space of index n_1 by a thick lens of index n. Prove that the lateral magnification is given by

$$\frac{n_1}{n_2} \cdot \frac{f}{U}$$

where f is the focal length of the lens and U is the object distance measured from the first principal focus of the lens.

36. Prove that, if an object is located inside a transparent sphere of radius R and index n at a distance R/n from the center, the lens is free from spherical aberration. Derive a formula for the position of the image. Is it real or virtual?

37. Discuss the axial spherical aberration of a convex mirror for a wide beam of parallel rays which are parallel to the mirror axis, computing the length of that part of the axis which is crossed by the reflected rays.

38. *a*. If the near point of an eye is 100 cm., what sort of lenses are needed in spectacles for reading? What focal length should these lenses have?

b. If spectacles with diverging lenses of focal length of 40 cm. are prescribed to correct defective vision, what is the far point for the eye?

39. A magnifying glass has a focal length of 4 cm. How far from the lens should an object be placed so that it is seen clearly by an observer whose eye is accommodated for a distance of 25 cm.? What is the magnifying power of the glass?

40. What is the magnifying power of a glass ball 2 cm. in diameter, if the index of refraction of the glass is 1.5?

41. A compound microscope comes equipped with objectives of focal lengths 2 and 6 mm. and with oculars of magnifying power 4× and 10×. Find the possible magnifying powers obtainable. What are the focal lengths of the oculars?

42. A properly focused telescope is sighted at the sun. How far and in what direction must the eyepiece be moved to project a sharp image of the sun on a screen 2 meters back of the eyepiece? The focal length of the eyepiece is 5 cm.

43. The objective of a field glass has a focal length of 24 cm. When used to examine an object 2 meters away, its magnifying power is 2.5. What is the focal length of the ocular? What will the magnifying power be when viewing an object at infinity?

44. The objective of an astronomical telescope has a focal length of 40 cm. and the ocular a focal length of 5 cm. Plot the magnifying power as a function of object distance if the latter varies from 5 meters to infinity. Through what distance must the ocular be moved to maintain a sharp focus for this variation of object distance?

45. What focal-length projecting lens is required to enlarge a 3- by 4-in. lantern slide to a 3- by 4-ft. image on a screen 25 ft. from the lens? Where should the slide be placed?

CHAPTER XV

DISPERSION AND SCATTERING

In this chapter we shall examine the question of the variation of the velocity of sinusoidal electromagnetic waves in dielectrics with the frequency of these waves. In the realm of optics, this effect is called the dispersion of light and makes itself evident in the variation of refractive index (and hence of the velocity of light) of nonconducting media with wave length or color. This study, when carried out in terms of an atomic picture, will lead us to an understanding of another phenomenon, the so-called scattering of light, as well as to a deeper insight into the nature of the refracted light wave, both in isotropic and in crystalline media. In Chap. XI we have given an atomic explanation of the electrostatic behavior of dielectrics, and we must now extend the ideas presented there to the case of electromagnetic fields varying with time. We have seen that the polarization of a dielectric by an electrostatic field resulted either from the orientation of permanent molecular dipole moments or from the atomic dipole moments induced by the field. Let us now inquire as to how we would expect these effects to vary as the frequency of an alternating field impressed on the dielectric is increased. For both types of polarization we may justifiably introduce the idea of a natural frequency (or frequencies). For the permanent dipoles the effect of temperature motion is to tend to restore a random orientation, very much as the restoring torque of a torsion pendulum tends to restore the equilibrium orientation of the pendulum. For the induced dipoles the restoring force, as we have seen in Sec. 52, may be taken as a linear restoring force, as if we had springs holding the electrons to the atoms.

In both the above cases the natural frequencies will depend on the stiffness coefficients of the force (or torque) and on the masses involved. In the case of the permanent dipoles the masses are extremely large compared to the electronic mass, and

consequently the natural frequencies for these dipoles are very much smaller than for the elastically bound electrons. For frequencies low compared to the natural frequencies, the steady-state motions of the permanent dipoles or of the electrons will be sinusoidal with amplitudes practically equal to the static displacements, as one always obtains when a system is set into forced oscillation by an external force of frequency very much below its natural frequency. As the frequency increases, the induced moments increase slowly until we reach frequencies not far from the lowest natural frequency. Here we encounter the phenomenon of resonance, a sharp increase of the amplitude of the forced motion with friction alone limiting the motion. In this range we have absorption of energy from the electromagnetic field, this absorbed energy heats the substance, and we say that we have an *absorption band* in the neighborhood of this natural frequency. The dielectric constant behaves "anomalously" in this region of frequencies, and, if these frequencies lie in the optical region, the substance is no longer transparent, the energy of the electromagnetic field being continually dissipated. Let us increase the frequency still more, passing well beyond the natural frequencies of the permanent dipoles. In this case the permanent dipoles no longer contribute to the dielectric constant (or index of refraction) of the body, since they can no longer follow the rapid variations of the field. For a substance like water, for which the abnormally high static dielectric constant of 80 is due to the permanent dipole moment of the water molecule, the permanent dipoles are completely ineffective at optical frequencies, with the result that the dielectric constant is much the same as that of a nonpolar substance at these frequencies.

Proceeding farther in the direction of increasing frequency, we must get into the infrared, visible, or ultraviolet region of the spectrum before the natural frequencies of the electrons which are responsible for the *induced* dipole moments become evident. We then have resonance phenomena, strong absorption of the electromagnetic waves at one or more frequencies, and regions of "normal" transparent behavior of the substance between these absorption frequencies. This is the region of the electromagnetic spectrum which concerns us in the study of optics, and we can entirely disregard the effects of the per-

manent dipoles in this connection. Our problem is the study of the motion of bound electrons under the influence of a sinusoidal electromagnetic wave. From the motion we can calculate the induced dipole moment as a function of the time, from this the polarization vector and the dielectric constant, leading us to a theory of the variation of index of refraction with frequency or wave length.

75. Dispersion in Gases.—In the case of a gas, in which the atoms are relatively far enough apart at moderate pressures so that we may neglect the interactions between particles, the calculation outlined above may be readily carried out. In the case of solids and liquids, the resultant force acting on an electron is the sum of the external force and those due to the neighboring induced dipoles; hence the calculation becomes more delicate. If an electromagnetic wave travels in a gas, the electrons in the atoms are set into an induced motion, and the resultant force acting on an electron is given by $e\,[\mathcal{E} + (v \times B)]$, where e is the charge on the electron, v its velocity, and $\mathcal{E}$ and B are the electric intensity and the magnetic induction vectors of the electromagnetic wave at the point where the electron is located. Since the electron velocities are very small compared to the velocity of light in the case at hand, we may neglect the magnetic force because it is of the order of magnitude of v/c times the electric force.

For the sake of simplicity, let us consider a plane electromagnetic wave, linearly polarized so that its electric vector has only an x-component, traveling in the z-direction. We have

$$\mathcal{E}_x = \mathcal{E}_0 \sin 2\pi\nu\left(t - \frac{z}{v}\right) \tag{1}$$

and, if we consider an atom located at a position $z = 0$, let us say, the force acting on an electron in this atom will be given by

$$F_x = e\mathcal{E}_x = e\mathcal{E}_0 \sin 2\pi\nu t \tag{2}$$

Here we have assumed that the phase of the electromagnetic wave does not change (at a given instant of time) over the region of space occupied by the atom. Since the diameter of an atom is about 10^{-8} cm. and the wave length of light about 5×10^{-5} cm. in the visible, we see that this is a justifiable procedure. It would be wrong, however, to proceed as above for the case of

X rays. Let us suppose that we are not in the immediate neighborhood of an absorption band (the natural frequency of the electron) and can therefore neglect "friction" effects. If the electron is displaced a distance x from its normal position, there will be a restoring force $-kx$ acting on it in addition to the force given by Eq. (2), and its equation of motion is

$$m\frac{d^2x}{dt^2} + kx = e\mathcal{E}_0 \sin 2\pi\nu t \tag{3}$$

This is the equation of the forced motion of a simple harmonic oscillator, and we are interested only in the *steady-state* motion which ensues with the same frequency as that of the external force. To find this motion, let us try a solution of the form[1]

$$x = A \sin 2\pi\nu t \tag{4}$$

being prepared to reject it if it fails, *i.e.*, if A does not turn out to be independent of t. From Eq. (4) we find for the acceleration

$$\frac{d^2x}{dt^2} = -4\pi^2\nu^2 A \sin 2\pi\nu t \tag{5}$$

and substituting the values of x and d^2x/dt^2, as given by Eqs. (4) and (5), back into Eq. (3), we find readily

$$A = \frac{e\mathcal{E}_0/4\pi^2 m}{\nu_0^2 - \nu^2} \tag{6}$$

where $\nu_0 = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$ is the natural frequency of the electron. Thus the required solution is

$$x = \frac{e\mathcal{E}_0 \sin 2\pi\nu t}{4\pi^2 m(\nu_0^2 - \nu^2)} = \frac{e\mathcal{E}_x/4\pi^2 m}{\nu_0^2 - \nu^2} \tag{7}$$

The dipole moment of the atom, due to the motion of this electron, is thus

$$ex = \frac{e^2\mathcal{E}_x/4\pi^2 m}{\nu_0^2 - \nu^2} \tag{8}$$

and the total induced dipole moment of the atom p_x is given by

$$p_x = \sum ex = \left(\sum \frac{e^2/4\pi m^2}{\nu_0^2 - \nu^2}\right)\mathcal{E}_x \tag{9}$$

[1] Compare the following discussion with that given in Frank's "Introduction to Mechanics and Heat," 2d ed., pp. 116–119.

where the summation is over all the electrons in the atom, different electrons having different natural frequencies. For the sake of simplicity we shall consider only one electron contributing appreciably to the induced dipole moment and have for the polarizability of the atom [compare Eq. (16), Chap. XI]

$$\alpha = \frac{p_x}{\mathcal{E}_x} = \frac{e^2}{4\pi^2 m(\nu_0^2 - \nu^2)} \tag{10}$$

Note that, for electrostatic fields ($\nu = 0$), this is the same result as is expressed in Eq. (16), Chap. XI. If there are N atoms per unit volume (or, far more precisely, N electrons of natural frequency ν_0 per unit volume), the polarization vector P has the magnitude

$$P = N\alpha\mathcal{E}$$

so that

$$D = \epsilon\mathcal{E} = \epsilon_0\mathcal{E} + 4\pi N\alpha\mathcal{E}$$

from which

$$\frac{\epsilon}{\epsilon_0} = \kappa = n^2 = 1 + \frac{4\pi N\alpha}{\epsilon_0} \tag{11}$$

[compare Eq. (17), Chap. XI]. Using the value of α as given by Eq. (10), there follows for the dielectric constant κ and the index of refraction n,

$$\kappa = n^2 = 1 + \frac{Ne^2/\pi m}{\nu_0^2 - \nu^2} \tag{12}$$

which predicts the variation of index of refraction with frequency ν, when only one type of electron plays an essential part. For the more general case, one has a sum of terms each similar to the right-hand term of Eq. (12), instead of the single term. For the case of gases, n is very nearly equal to unity (for air, $n = 1.0003$), so that we may write instead of Eq. (12)

$$n - 1 = \frac{Ne^2/2\pi m}{\nu_0^2 - \nu^2} \tag{13}$$

In Fig. 193 is shown a plot of Eq. (13). According to our equation $n - 1$ becomes infinite for $\nu = \nu_0$, but, as we have stated, in this region strong absorption takes place; consequently our equation is essentially correct for frequencies less than ν_A and greater than ν_B as indicated in the figure, but not in between. Notice that, in the range for which Eq. (13) is expected to hold,

the index of refraction *increases with increasing frequency, i.e.,* with decreasing wave length. This is the so-called *normal dispersion,* whereas the behavior in the neighborhood of $\nu = \nu_0$ is called "anomalous," since it can be shown that the refractive index decreases with increasing frequency in this region. It is customary for experimentalists to write formulas for dispersion in terms of wave length rather than frequency. One of the

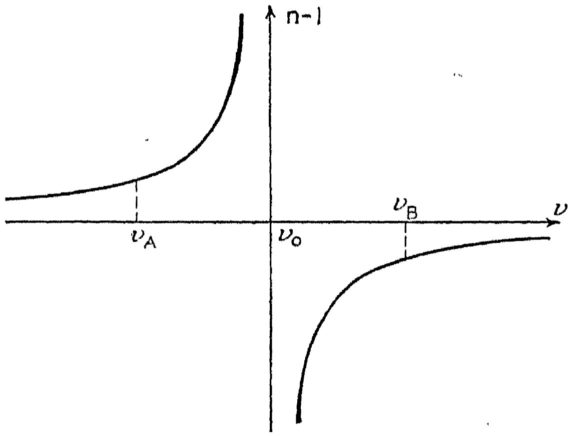

FIG. 193.

common empirical formulas, known as the Cauchy formula, is as follows:

$$n = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} \tag{14}$$

One can easily show that Eq. (13) or (12) leads to this form for $\nu << \nu_0$, *i.e.,* when the absorption is in the ultraviolet region of the spectrum.

76. Dispersion in Solids and Liquids; The Prism Spectroscope. The theory of dispersion in solids and liquids proceeds along the lines of that given for gases, with the exception that Eq. (2) for the force on an electron must be replaced by an appropriate expression involving the effect of the neighboring atoms. We shall not enter into this question but wish to point out that the empirical formula (14), which works well for ordinary optical materials in the visible region of the spectrum, can be justified on theoretical grounds. In Fig. 194 is shown the variation of index of refraction with wave length for two common types of optical glass, both showing normal dispersion. As a general rule, the rate of increase of n with decreasing wave length is

larger at short wave lengths, and for different substances, the curve is usually steeper at a given wave length, the larger the value of n.

The dispersion of optical materials is utilized in analyzing the spectral composition of light in the instrument known

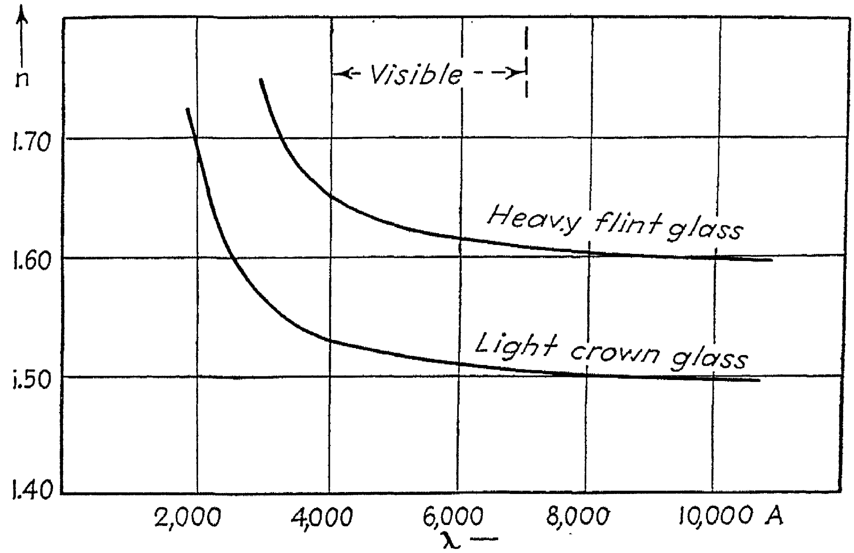

FIG. 194.

as the *prism spectrometer*. A beam of light, in general, consists of an electromagnetic wave field which is the superposition of many elementary waves of different frequencies (colors), amplitudes and polarization, and the study of the wave lengths of the various component waves is fundamental in obtaining information concerning the structure of atoms and molecules. We start with the problem of the deviation of a very narrow pencil of light rays by a prism of angle A, as shown in Fig. 195. Let us suppose that we have monochromatic light, *i.e.*, light of a single color. For a fixed prism angle A, the angle of deviation δ varies with the angle of incidence i_1, and there is one angle of incidence for which the deviation is a minimum. From the figure it is evident that

$$\delta = i_1 - r_1 + r_2 - i_2 = i_1 + r_2 - A \tag{15}$$

If the deviation δ is to be a minimum, we must have $d\delta/di_1 = 0$, or from Eq. (15)

$$\frac{dr_2}{di_1} = -1 \tag{16}$$

so that we must express r_2 in terms of i_1. This is most easily done in differential form and suffices for the condition imposed by Eq. (16). At each surface of the prism, we have from the law of refraction

$$\left.\begin{aligned} \sin i_1 &= n \sin r_1 \\ n \sin i_2 &= \sin r_2 \end{aligned}\right\} \tag{17}$$

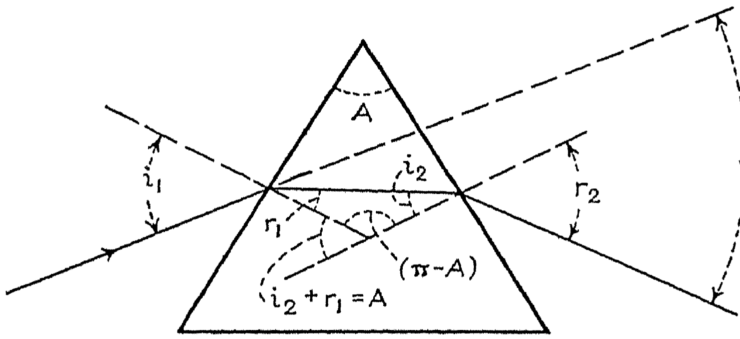

FIG. 195.

Differentiating these expressions, we obtain

$$\begin{aligned} \cos i_1 \, di_1 &= n \cos r_1 \, dr_1 \\ n \cos i_2 \, di_2 &= \cos r_2 \, dr_2 \end{aligned}$$

and since $r_1 + i_2 = A$, we have the relation $dr_1 = -di_2$. Thus there follows

$$dr_2 = n \frac{\cos i_2}{\cos r_2} di_2 = -n \frac{\cos i_2}{\cos r_2} dr_1 = -\frac{\cos i_2}{\cos r_2} \cdot \frac{\cos i_1}{\cos r_1} di_1$$

or, rewritten,

$$\frac{dr_2}{di_1} = -\frac{\cos i_1}{\cos r_2} \cdot \frac{\cos i_2}{\cos r_1}$$

and Eq. (16) requires that

$$\frac{\cos i_1}{\cos r_2} \cdot \frac{\cos (A - r_1)}{\cos r_1} = 1 \tag{18}$$

where we have used the relation $r_1 + i_2 = A$. Equation (18) is evidently satisfied if $i_1 = r_2$ and $r_1 = A/2$, *i.e.*, when the ray passes symmetrically through the prism. When this condition is satisfied, we have from Eq. (15)

$$i_1 = \frac{A + \delta}{2}$$

so that the first of Eqs. (17) gives the condition

$$\sin\left(\frac{A+\delta}{2}\right) = n \sin\frac{A}{2} \tag{19}$$

giving the *minimum deviation* δ in terms of the prism angle A and the refractive index n. For small angle prisms, we may replace the sines of the angles in Eq. (19) by the angles, so that Eq. (19) becomes

$$n = A + \delta$$

or

$$\delta = (n - 1)A \tag{20}$$

which is a convenient approximation. Note that the deviation increases with increasing refractive index, so that the shorter the wave length, the greater the deviation for a given prism. If the light incident on the prism is white light, a mixture of all visible wave lengths, the emergent light will be spread or "dispersed" into a spectrum, as shown in Fig. 196. One defines the *angular dispersion* D as the rate of change of the deviation with wave length, *i.e.*,

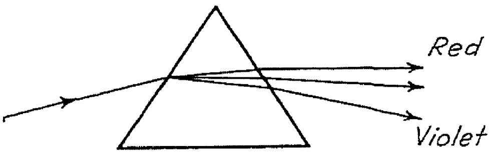

FIG. 196.

$$D \quad \frac{d\delta}{d\lambda} = A\frac{dn}{d\lambda}$$

and last equality holding for small angle prisms. The actual dispersion of a prism spectroscope (see below) is measured by the separation in angstroms per millimeter in the field of the telescope or on the photograph of the spectrum. This quantity depends not only on the angular dispersion D but also on the focal length of the telescope or camera objective.

It is customary to define the *dispersive power* d of a small angle prism as the ratio of the difference in deviation for two extreme colors in the visible spectrum to the mean deviation of the spectrum as a whole. In this connection, it has become common practice to utilize the following wave lengths as reference wave lengths:

$\lambda_1 = 6{,}560$ A. (the Fraunhofer C-line; red)
$\lambda_2 = 4{,}860$ A. (the Fraunhofer F-line; blue)
$\lambda_0 = 5{,}890$ A. (the Fraunhofer D-line; yellow)

Using Eq. (20), we find immediately for the dispersive power of a small angle prism

$$d = \frac{\delta_2 - \delta_1}{\delta_0} = \frac{n_2 - n_1}{n_0 - 1} \tag{21}$$

where the n's are the refractive indices for the above colors. The refractive indices and dispersive powers are tabulated for two kinds of glass:

	n_0	n_1		
Silicate flint glass....	1.620	1.613	1.632	0.031
Silicate crown glass...	1.508	1.504	1.513	0.018

Since, in general, the dispersive power of various optical materials is not proportional to the mean refractive index, it is

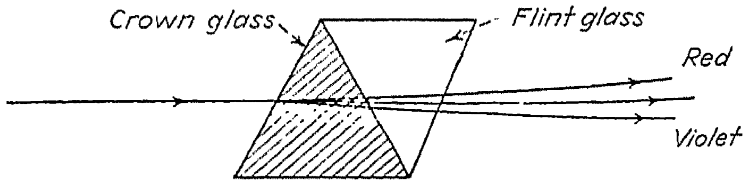

FIG. 197.

possible to combine two or more prisms to obtain no net deviation of a given wave length, such as the D-line, and yet obtain a spectrum. Such a device is known as a *direct-vision prism* (Fig. 197) and is used in the direct-vision spectroscope, a convenient small instrument which may be held in the hand and pointed directly at the light source.

It is also possible to combine two prisms so that the dispersions are compensated (strictly speaking, for two wave lengths only), but the deviations are not. Such a prism is called "achromatic." This principle is utilized in designing achromatic lenses, and we shall illustrate the method in the next section.

The ordinary form of the prism spectroscope is shown in Fig. 198. The essential parts are as follows: A narrow slit, S, illumi-

nated by the light to be examined; a collimating lens C which produces a parallel beam of light incident on the prism P; a telescope lens T which produces real images of the slit at the first focal point of the eyepiece E. The observer sees a series of images of the slit side by side, each of different color. If white light is used, the images overlap and a *continuous spectrum* is observed. If the incident light is a mixture of a finite number of wave lengths, a series of bright lines, one of each color present in the incident light, will appear and a *line spectrum* is observed.

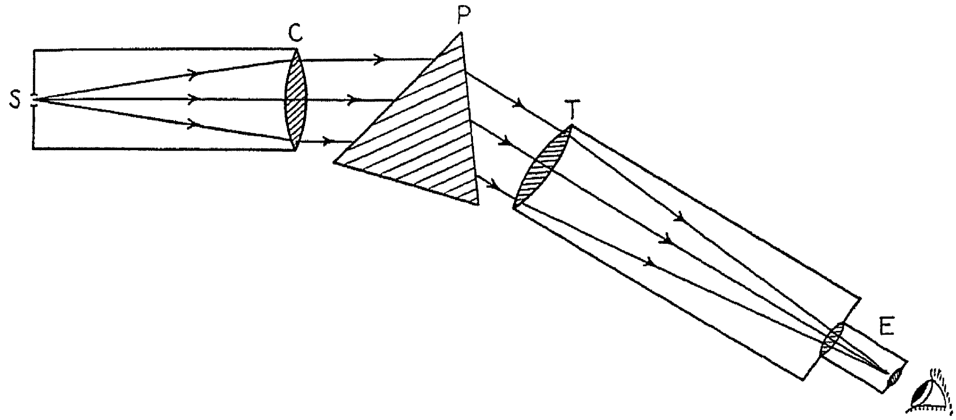

FIG. 198.

A stigmatic pencil incident on a prism becomes astigmatic upon emergence unless the angle of incidence is that for minimum deviation, so that the collimating lens is necessary to produce sharp images. This lens must be achromatic (corrected for chromatic aberration).

77. Chromatic Aberrations and Their Correction.—The phenomenon of dispersion results in the chromatic aberration of simple lenses, as mentioned in the last chapter. Since the focal length of a lens depends on its index of refraction, a single lens will produce series of colored images of an object at different positions and of different magnifications. It is possible to design a lens which will have the same focal length for two wave lengths. It is then said to be "achromatized" for these two colors. The following method illustrates how an "achromatic doublet," consisting of two thin lenses of different kinds of glass in contact, may be designed to be achromatic for the C and F Fraunhofer lines mentioned in the previous section. For one of these lenses we have for any single wave length

$$\frac{1}{f'} = (n' - 1)\left(\frac{1}{R'_1} - \frac{1}{R'_2}\right) = c'(n' - 1)$$

and for the other

$$\frac{1}{f''} = (n'' - 1)\left(\frac{1}{R''_1} - \frac{1}{R''_2}\right) = c''(n'' - 1)$$

The focal length f of the two lenses is related to f' and f'' by

$$\frac{1}{f} = \frac{1}{f'} + \frac{1}{f''} = c'(n' - 1) + c''(n'' - 1) \tag{23}$$

and, if f is to be the same for both the C- and F-lines, we must have

$$c'(n'_1 - 1) + c''(n''_1 - 1) = c'(n'_2 - 1) + c''(n''_2 - 1)$$

or

$$\frac{c'}{c''} = -\frac{n''_2 - n''_1}{n'_2 - n'_1} \tag{24}$$

The focal lengths of the two lenses for the intermediate D-line are

$$\frac{1}{f'_0} = (n'_0 - 1)c' \qquad \text{and} \qquad \frac{1}{f''_0} = (n''_0 - 1)c''$$

so that

$$\frac{c'}{c''} = \frac{n''_0 - 1}{n'_0 - 1} \cdot \frac{f''_0}{f'_0} \tag{25}$$

Combining Eqs. (24) and (25) one finds readily, using Eq. (21),

$$\frac{f''_0}{f'_0} = -\frac{d''}{d'} \tag{26}$$

where d'' and d' are the dispersive powers of the two kinds of glass utilized. Since d' and d'' are positive, the doublet must consist of a positive and a negative lens. If the faces of the lenses in contact are to be cemented together, we must have $R'_2 = R''_1$. We now have three conditions imposed on the four radii of curvature.

$$R'_2 = R''_1$$

$$\frac{f''_0}{f'_0} = -\frac{d''}{d'} = \frac{(n'_0 - 1)\left(\frac{1}{R'_1} - \frac{1}{R'_2}\right)}{(n''_0 - 1)\left(\frac{1}{R''_1} - \frac{1}{R''_2}\right)} \tag{27}$$

and

$$\frac{1}{f} = (n'_2 - 1)\left(\frac{1}{R'_1} - \frac{1}{R'_2}\right) + (n''_2 - 1)\left(\frac{1}{R''_1} - \frac{1}{R''_2}\right)$$

A fourth condition may be imposed arbitrarily and may be utilized for other corrections such as minimizing spherical aberration. For simplicity, if we choose $R = \infty$ making the first lens plano-convex with the plane side toward the light source, we obtain the lens illustrated in Fig. 199. Numerical examples are left to the problems.

It is also possible to achromatize for focal lengths by constructing a compound lens of two simple lenses of the same kind of glass separated from each other by an appropriate dis-

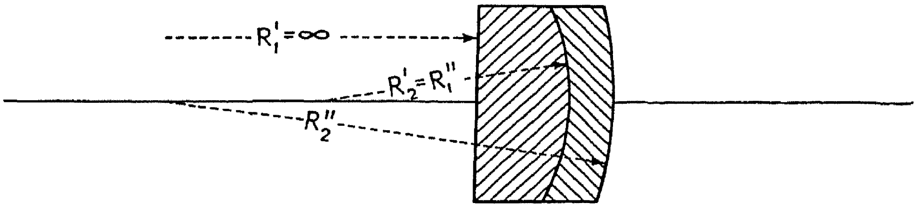

FIG. 199.

tance a. The equivalent focal length of the combination, f, is given by

$$\frac{1}{f} = \frac{1}{f'} + \frac{1}{f''} - \frac{a}{f'f''} \tag{28}$$

where f' and f'' are the focal lengths of the two lenses employed. Since

$$\frac{1}{f'} = (n - 1)\left(\frac{1}{R'_1} - \frac{1}{R'_2}\right) = c'(n - 1)$$

and

$$\frac{1}{f''} = c''(n - 1)$$

we have

$$\frac{1}{f} = (c' + c'')(n - 1) - ac'c''(n - 1)^2 \tag{29}$$

If f is to be independent of wave length, we must have $\dfrac{d}{d\lambda}\left(\dfrac{1}{f}\right) = 0.$ This yields

$$(c' + c'')\frac{dn}{d\lambda} - 2ac'c''(n - 1)\frac{dn}{d\lambda} = 0$$

or

$$\frac{dn}{d\lambda}[c' + c'' - 2ac'c''(n - 1)] = 0$$

Since $dn/d\lambda \neq 0$, we must have

$$a = \frac{c' + c''}{2c'c''\,(n - 1)} \doteq \frac{(1/f') + (1/f'')}{2/f'f''} = \frac{f' + f''}{2} \tag{30}$$

so that the separation of the lenses should be equal to one-half the sum of their focal lengths. The correction can be made exactly only for one wave length, namely, for that wave length to which f' and f'' refer. However, the departure from perfect achromatization is small for wave lengths not far from the one for which Eq. (30) is satisfied, and this method is employed in many types of oculars.

78. The Scattering of Light.—In our calculation of the index of refraction of a gas as a function of frequency in Sec. 75, we were concerned solely with the magnitudes of the dipole moments induced by a light wave traveling through the gas. We must now consider the fact that the induced oscillating dipoles are themselves radiators of electromagnetic waves, and these secondary waves are responsible for the so-called *scattering* of light by matter. Suppose a linearly polarized monochromatic wave is incident on an electron bound to an atom of a gas. The electron performs forced simple harmonic motion and gives rise to an oscillating dipole moment given by Eq. (8), *viz.*,

$$p = ex = \frac{e^2\mathcal{E}_0}{4\pi^2 m}\frac{\sin 2\pi\nu t}{(\nu_0^2 - \nu^2)}$$

where e and m are the electronic charge and mass, respectively, $\mathcal{E}_0$ is the amplitude of the electric vector of the incident wave and ν_0 is the natural frequency of the oscillating dipole. Absorption is neglected as before. This oscillating dipole will radiate according to Eq. (19), Chap. IX, and the light emitted by all the induced dipoles constitutes the scattered light. The amplitude P of the oscillating dipole is evidently

$$P = \frac{e^2\mathcal{E}_0}{4\pi^2 m(\nu_0^2 - \nu^2)} \tag{31}$$

so that the average rate of emission of energy of this electron is, using Eq. (19) of Chap. IX,

$$\frac{e^4\mathcal{E}_0^2}{3m^2c^3}\frac{1}{\left(\frac{\nu_0^2}{\nu} - 1\right)^2} \tag{32}$$

Owing to the random thermal motion of the atoms of a gas and the consequent random positions of these atoms, there will be no fixed relations among the phases of the waves emitted from the various atoms—with the important exception of the light emitted *in* the direction of the incident wave—so that the intensity of the scattered light will be the sum of the intensity contributions from the individual dipoles. If there are N dipoles per unit volume, the intensity of the scattered light will be N times the expression (32). As a measure of the scattering power, we take the ratio of the light scattered per unit volume to the intensity of the incident light wave. Since the latter has the average value $c\mathcal{E}_0^2/8\pi$, we find for the scattering power,

$$\text{Scattering power} = \frac{8\pi Ne^4}{3m^2c^4} \frac{1}{\left(\frac{\nu_0^2}{\nu^2} - 1\right)^2} \tag{33}$$

If the natural frequency ν_0 lies in the ultraviolet, as it does for a normal atom, then for visible light we may set $\nu_0^2/\nu^2 >> 1$, so that Eq. (33) takes the form

$$\text{Scattering power} = \frac{8\pi Ne^4\nu^4}{3m^2c^4\nu_0^4} = \frac{8\pi Ne^4}{3m^2\nu_0^4\lambda^4} \tag{34}$$

where λ is the wave length. This is the so-called Rayleigh scattering formula which predicts that blue light should be scattered much more than red light because of the inverse fourth power dependence on wave length. This provides an explanation of the blue color of the sky, the air molecules scattering the blue light much more than the red. The transmitted light is correspondingly deficient in blue light as is evidenced in the red color of the sun at sunset, when the sunlight reaches us through a long air path.

On the other hand, for high frequencies for which $\nu^2 >> \nu_0^2$, Eq. (33) takes the form

$$\text{Scattering power} = \frac{8\pi Ne^4}{3m^2c^4} \tag{35}$$

independent of wave length. This is the Thomson scattering formula, and it finds important application in the field of X rays. Finally, if ν is almost equal to ν_0, the scattered light becomes very intense, and then we may no longer neglect absorption

("frictional effects"). This scattering is called *resonance scattering*.

Since all the induced dipoles vibrate in the direction of the electric vector $\mathcal{E}_0$ of the incident light, the scattered light will be linearly polarized if the incident light is so polarized. There will be zero intensity of scattered light in the direction of the vector $\mathcal{E}_0$ and a maximum at right angles to this direction, both directions being normal to the direction of propagation of the incident wave. If the incident light is unpolarized, the light scattered in a direction perpendicular to the direction of propagation of the incident light will be linearly polarized, only the components of the electric vectors of the elementary waves, of which unpolarized light is composed, which are normal to both the above directions being effective in producing scattered light in this given direction. The fundamental experiments performed by *Barkla*, which showed that X rays are electromagnetic waves, depended on these facts.

We have already mentioned the fact that the atoms of a substance scatter independently of one another (so that one may add intensities rather than amplitudes), because of thermal agitation, in all directions except that of the incident wave. In this forward direction the scattered light combines with the incident light to form the refracted light wave. It is evident that the relative phase of the incident wave and the wave scattered by a given atom at a point lying ahead of this atom does *not* depend on the exact position of the scattering atom, so that there are *fixed* phase relations between the incident wave and the waves scattered in the forward direction. Thus we must add amplitudes rather than intensities, and interference effects become important. Owing to the phase retardations of the secondary scattered waves relative to the primary incident wave, the resultant refracted wave suffers a phase retardation relative to the primary wave proportional to the distance traversed, *i.e.*, to the number of scattering atoms in this path, with the result that this resultant wave progresses with a phase velocity less than that of electromagnetic waves in empty space. This yields a simple and effective picture of the manner in which the refracted wave is produced and of the reason for the reduced velocity of light in material media.

To clarify the preceding statements, let us consider a simple model in which we imagine the atoms arranged in layers perpendicular to the direction of propagation of the incident plane wave *WW* in Fig. 200. This incident wave excites the electrons in the layer *AA*, and they emit a secondary plane wave out of phase with the primary wave. We can find the phase of this secondary wave by a simple vector diagram. At the point *B* the electric vector oscillates sinusoidally with time, so that we can use the vector diagram of a.c. circuits. Furthermore, the amplitude of the secondary wave emitted by a single layer of atoms is exceedingly small compared to that of the primary wave, and,

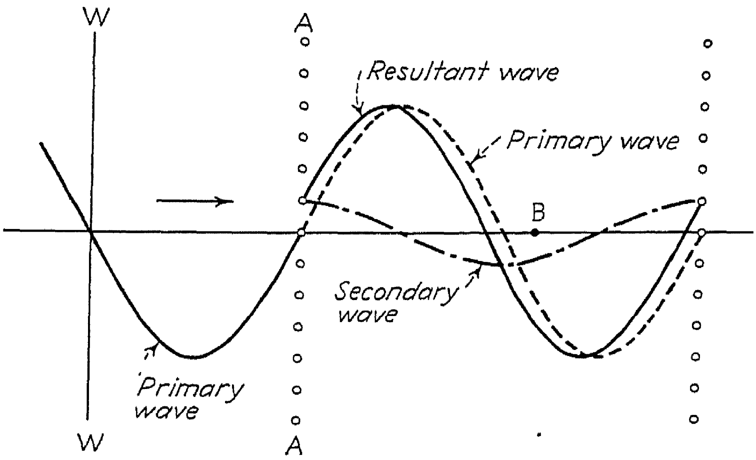

FIG. 200.

since there is no absorption, the amplitude of the resultant wave must equal the amplitude of the incident wave. In Fig. 201 $\mathcal{E}_p$ represents the amplitude of the primary wave at the point *B*, assuming no layer of atoms. $\mathcal{E}_s$ is the amplitude of the secondary wave, which we take as infinitesimal, and the sum of these yields $\mathcal{E}_R$, the amplitude of the resultant wave. Since $\mathcal{E}_p$ and $\mathcal{E}_R$ are equal in magnitude, the vector $\mathcal{E}_s$ must be practically at right angles to either, so that the secondary wave is very nearly 90° out of phase with the primary wave. This does not contradict the fact that the secondary wave emitted by a single atom is 180° out of phase with the primary wave (the acceleration being 180° out of phase with the displacement of the electron), since we are considering the contribution at *B* of all the atoms in the layer *AA*, and the individual secondary waves arriving at *B* from the individual atoms have different

amplitudes and phases. The sum of these individual spherical waves then yields the plane secondary wave of the phase which we have calculated. In Fig. 200 are shown the relations between primary, secondary, and resultant waves. The phase shift δ shown in Fig. 201, produced by *one* layer of atoms, is proportional to the amplitude of the secondary wave, and, as the wave progresses through the medium, the total phase shift in a distance x will be proportional to the number of atomic layers in this distance and hence to the distance x for a homogeneous substance.

Summarizing, we have the result that the refracted wave in a material medium is the superposition of the incident wave and the light scattered in the forward direction, whereas in other directions we have the ordinary scattered light as expressed in Eq. (33).

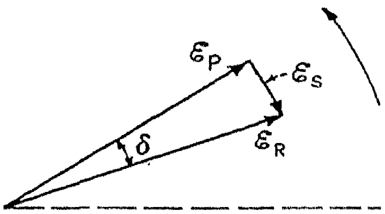

FIG. 201.

79. The Propagation of Light in Crystals; Double Refraction.—In isotropic media the velocity of light is independent of the direction of propagation and of the polarization of the light waves. According to the theory set forth in this chapter, this fact indicates that the natural frequencies ν_0 of the electrons which are set into oscillation in such media are independent of the direction in which these oscillations occur. The dependence of refractive index on these natural frequencies is given essentially by the right-hand term in Eq. (12) (the equation itself must be modified for the case of solids), and in the region of normal dispersion the velocity of light should increase with increasing ν_0, *i.e.*, the stronger the binding forces acting on the electrons. There exist many transparent crystals which are *anisotropic*, the binding forces and natural frequencies ν_0 of the electrons being different in different directions, and a number of remarkable and important optical phenomena occur when light passes through such crystals.

We shall confine our attention to the case for which there is *one* preferred direction in the crystal, let us say the x-axis, so that all directions in planes normal to this x-axis (the y-z planes) are equivalent. Thus, if the electrons are set in motion along the x-axis, they will behave as simple harmonic oscillators of natural frequency ν_{ox}, whereas, if they are set into motion in

any direction in a y-z plane, they behave as harmonic oscillators of a different natural frequency $\nu_{oy} = \nu_{oz}$. Crystals for which there is *one* preferred direction are called *uniaxial*, and this preferred direction is called the *optic axis* of the crystal. Consider a plane linearly polarized light wave traveling in the direction of the optic axis, *i.e.*, along the x-axis. Since the electric vector oscillates in the y-z plane, the amplitudes of the induced oscillating dipoles and consequently the velocity of propagation of the wave do not depend on the direction of polarization. Hence we may introduce an ordinary index of refraction n_0 (depending on ν_{oy}) for propagation along the optic axis and the velocity of propagation v_0 for this case is given by

$$v_0 = \frac{c}{n_0} \tag{36}$$

Thus if light is normally incident on a crystal surface which is perpendicular to the optic axis, it propagates through the crystal with a velocity v_0 independent of the state of polarization of the light, and the crystal behaves like an isotropic substance, such as glass.

The state of affairs is quite different, however, if the direction of propagation is not along the optic axis. Consider a plane wave propagating in a direction perpendicular to the optic axis, let us say along the z-axis, such as one may obtain by allowing parallel light to fall at normal incidence on a crystal surface which has been ground parallel to the optic axis. In this case it is evident that the state of polarization of the wave plays an important role in determining the nature of the refracted wave. First, let us consider the case for which the incident wave is linearly polarized, the direction of oscillation of the electric vector being along the y-axis (the direction AA in Fig. 202) perpendicular to the direction of the optic axis. The electrons in the crystal are set into oscillation in the y-direction, and the phase velocity of this wave inside the crystal depends on the natural frequency ν_{oy} in accordance with the foregoing discussion. Evidently this wave will propagate with the velocity v_0 given by Eq. (36).

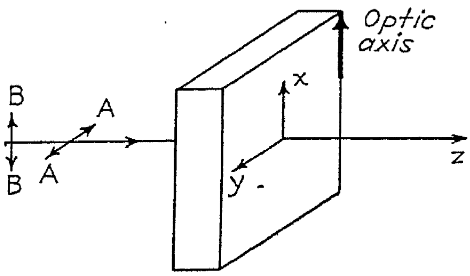

Fig. 202.

On the other hand, if the incident wave is polarized along the optic axis (the direction BB in Fig. 202), the electrons in the crystal are set into oscillation along the x-axis. Thus the phase velocity of this wave inside the crystal depends on the natural frequency ν_{ox}, which is different from ν_{oy}. We may now introduce a second index of refraction n_e for this type of wave, so that its velocity v_e is given by

$$v_e = \frac{c}{n_e}$$

n_0 and n_e are called the *principal indices of refraction* of the crystal. If $v_0 > v_e (n_0 < n_e)$, the crystal is called *positive uniaxial,* whereas, if $v_0 < v_e (n_0 > n_e)$, it is called *negative uniaxial.* For the crystal calcite ($CaCO_3$), a crystal important in many applications to optical instruments, the principal indices have the values (for the sodium D-lines)

$$n_0 = 1.6584$$
$$n_e = 1.4864$$

so that calcite is a negative uniaxial crystal.

Now let the incident light have an arbitrary direction of polarization relative to the optic axis, as indicated in Fig. 203. We can resolve the electric vector of the incident wave into x- and y-components (Fig. 203), and these oscillate in phase for the linearly polarized wave under discussion. [For elliptically (or circularly) polarized light we would have the same vector diagram as in Fig. 203, but the two components would then oscillate out of phase with each other.] After passing through the crystal the x- and y-components of $\mathcal{E}$ will no longer be in phase because of the different velocities of propagation for these two types of polarization. The phase difference thus obtained depends on the thickness of the crystal, and in general the emergent light will be polarized differently from the incident light. Suppose, for example, that the crystal of Fig. 202 is just thick enough to produce a phase difference of $\lambda/2$ (or 180°) between the x- and y-components of $\mathcal{E}$. For the emergent light

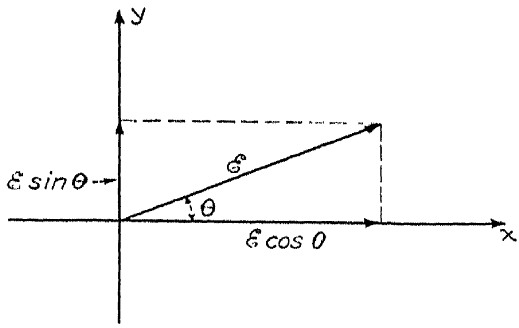

FIG. 203.

we then can construct the vector diagram corresponding to Fig. 203 for the incident light. This is shown in Fig. 204. The resultant $\mathcal{E}'$ of the emergent light now makes a negative angle θ with the optic axis. The effect of the crystal has been to rotate the direction of polarization through an angle of 2θ. We can readily derive a formula for the thickness of such a *half-wave*

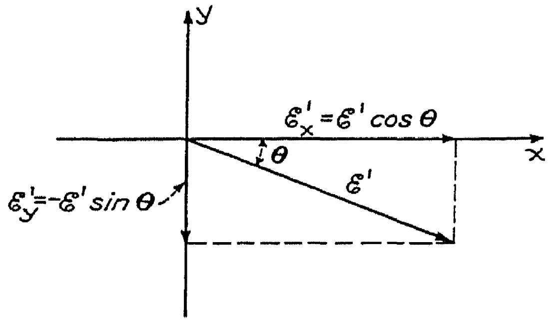

FIG. 204.

plate. The phase lag for the y-component of $\mathcal{E}$ produced by a plate of thickness d is given by

$$\frac{2\pi\nu d}{v_0} \qquad \frac{2\pi\nu d n_0}{c}$$

whereas for the x-component it is

$$2\pi\nu d \qquad 2\pi\nu d n_e$$

Remembering that c/ν is the wave length of the light (in vacuum), the relative phase retardation between these two components of $\mathcal{E}$ becomes

$$\delta = \frac{2\pi d}{\lambda}(n_0 - n_e) \tag{38}$$

For a half-wave plate we set $\delta \quad \pi$ and have from Eq. (38)

$$d = \frac{\lambda}{2}\,\frac{1}{n_0 - n_e}$$

Evidently the same results would be obtained for a given wave length if the phase retardation is an odd number of half wave lengths. In the special case of $\theta = 45°$ in Fig. 203, a quarter-wave plate will produce circularly polarized emergent light. Further details are left to the problems.

There are a number of uniaxial crystals, such as tourmaline, which, when used as in Fig. 202, have the remarkable property of absorbing light polarized along the y-axis and transmitting light polarized along the x-axis (the optic axis). Such crystals are called *dichroic*. A tourmaline plate used as indicated in Fig. 205 forms a simple polarizer. Incident unpolarized light falling on the plate emerges as linearly polarized light as indicated. Although the resultant electric vector of unpolarized light may be represented by two equal components at right angles, it must be remembered that, since the electric vector is the sum of many elementary vectors of random orientation and phases, there are no definite phase relations between these two components. The commercial material "polaroid," composed of many small

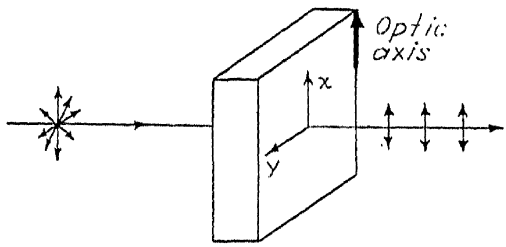

FIG. 205.

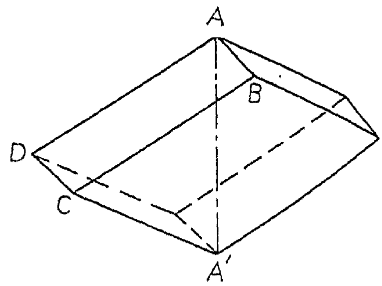

FIG. 206.

dichroic crystals which are lined up, acts in the above manner and provides a very useful and simple polarizer.

Finally, we must consider what happens when light is incident on a crystal surface which is neither parallel to nor perpendicular to the optic axis. For simplicity, we shall consider the case of normal incidence on one of the natural cleavage planes of calcite. The natural cleavage planes of calcite yield a rhombohedral crystal as shown in Fig. 206, in which AA' gives the direction of the optic axis when the crystal is of such a length BC that AA' makes equal angles with the three edges at A. If a narrow beam of natural unpolarized light falls at normal incidence on the face $ABCD$ of the crystal, one observes that *two* refracted beams are formed inside the crystal and two parallel beams emerge from the opposite face, and these two beams are linearly polarized at right angles to each other.

This phenomenon of *double refraction* or *birefringence* is illustrated in Fig. 207. The incident beam I is normally incident on the left-hand face at C. Inside the crystal it breaks up into

two beams, one of which, O, traverses the crystal without being bent and the other, E, is refracted upon entrance in the direction CD and emerges parallel to the first beam. The ray O which goes straight through the crystal behaves in the crystal, in this experiment and in any others which one may perform, just as if it were in an isotropic medium. It is called the *ordinary ray* and

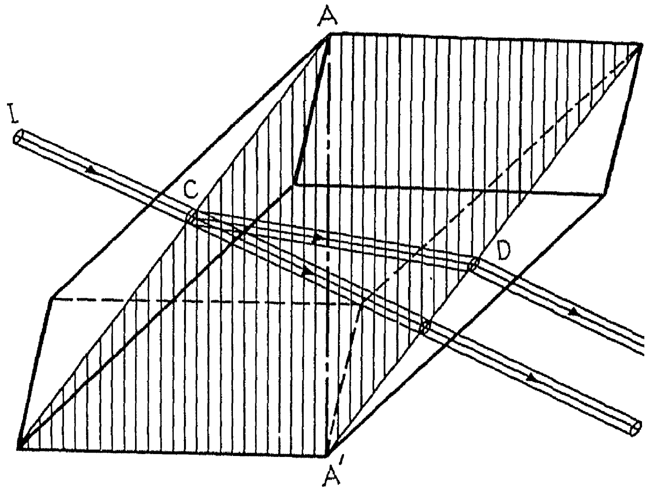

FIG. 207.

obeys Snell's law when the light is not normally incident on the surface. Its velocity is v_0 as given by Eq. (36) and is independent of its direction of propagation. On the other hand, the ray E evidently does not follow Snell's law, as it is refracted for normal incidence. It is called the *extraordinary ray*, and its velocity may have any value between v_0 and v_e as given by Eqs. (36) and (37). To answer the question as to the direction in which the extraordinary ray is refracted, one observes that if the crystal in Fig. 207 is rotated about ICO as an axis, the position of the ordinary ray O is unaltered but the ray CDE rotates about this axis. Hence CD is fixed in the crystal in the shaded plane which contains the optic axis and the incident ray. This plane is called a *principal section* of the crystal for this ray, and the extraordinary ray lies in this plane. Figure 208 indicates the process somewhat more simply than Fig. 207, and in it the plane of the page is chosen to

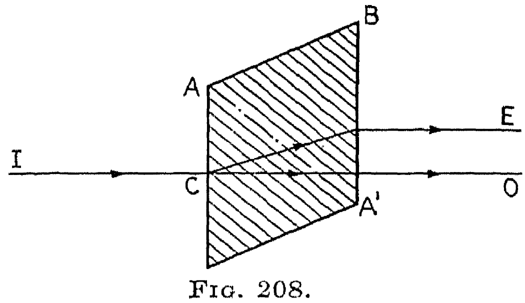

FIG. 208.

coincide with the principal section. The ordinary ray is polarized in a direction perpendicular to the principal section, *i.e.*, to the plane of the page, whereas the extraordinary ray is polarized in the plane of the page in this figure. For light not normally incident on the surface of a crystal, one obtains a very similar picture to the one which we have discussed, the ordinary ray obeying Snell's law with an index n_0 and the extraordinary ray behaving

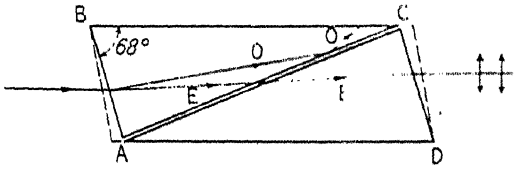

FIG. 209.

anomalously. Since a quantitative investigation of double refraction would lead us far beyond the scope of this book, we shall content ourselves with the above brief description.

The Nicol Prism.—A beam of linearly polarized light may be obtained conveniently with the help of a Nicol prism or "Nicol" which is constructed from a long calcite rhombohedron as follows: In the principal section of a calcite crystal (Fig. 209) the angles at B and D are 71°. The two end faces AB and CD are cut down so that these angles are reduced to 68°. The crystal is sliced along AC in a plane perpendicular to the ends and to the plane of the paper (the principal section), and the two surfaces thus formed are cemented together with Canada balsam of refractive index 1.55 which is less than that for the ordinary ray in calcite and greater than that for the extraordinary ray shown in the figure. The passage of a ray of light through the prism is shown in Fig. 209, the ordinary ray being totally reflected from the Canada balsam and the extraordinary ray being transmitted with but small loss in intensity. The emergent light is linearly polarized in the plane of the paper as indicated and the direction of oscillation is along the short diagonal of the diamond-shaped end face of the Nicol prism (Fig. 210).

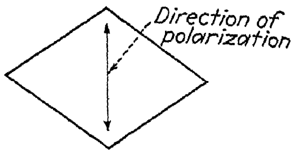

FIG. 210.

Problems

1. The observed values of $(n - 1)$ for hydrogen at a number of wave engths are as follows:

λ, angstroms...	5,460	4,080	3,340	2,890	2,540	2,300	1,900
$(n - 1) \times 10^7$.	1,400	1,426	1,461	1,499	1,547	1,594	1,718

Make a plot of these data, and from two points on the curve compute the values of N and ν_0 (or the corresponding wave length λ_0) of Eq. (13) of the text. How does the value of N compare with the number of hydrogen molecules per cubic centimeter under standard conditions?

Using these values of N and ν_0, plot $(n - 1)$ as a function of wave length according to Eq. (13), and compare with the experimental curve.

2. Show that the Cauchy dispersion formula [Eq. (14)] follows from Eq. (13) if $\nu << \nu_0$. Prove that the Cauchy formula [Eq. (14)] is true for a gas with more than one kind of electron per atom (different ν_0's) providing all the ν_0's are larger than ν.

3. Compute the prism angle for a crown-glass prism needed to construct a direct-vision prism in conjunction with an 8° flint-glass prism, if the Fraunhofer D-line is to have no net deviation.

Find the angular separation of the Fraunhofer C- and F-lines.

Use the data given in the text for the refractive indices of crown and flint glass.

4. What is the proper angle for the crown-glass prism of Prob. 3 if, when used with the 8° flint-glass prism, there is to be no net dispersion (angular separation) of the C- and F-lines?

Compute the net deviation of the F-, D-, and C-lines.

5. The following table gives the index of refraction as a function of wave length of two types of glass.

Wave length, angstroms	n (light crown glass)	n (heavy flint glass)
4,000	1.5238	1.8059
4,500	1.5180	1.7843
5,000	1.5139	1.7706
5,500	1.5108	1.7611
6,000	1.5085	1.7539
6,500	1.5067	1.7485
7,000	1.5051	1.7435
7,500	1.5040	1.7389

Using these data, compute the design of an achromatic doublet of focal length -20.0 cm., achromatized in focal length for wave lengths of 4,500 and 6,500 A. Take one of the lens surfaces as plane.

Compute and plot the focal length of the lens so designed as a function of wave length from 4,000 to 7,500 A.

6. Prove that two simple harmonic motions at right angles of arbitrary amplitudes and phases but of equal frequency yield elliptical motion. Under what conditions does this ellipse become a circle? A straight line?

7. Compute the thickness of a quarter-wave plate of calcite and of quartz for the wave length $\lambda = 5{,}893$ A. (the D line). The principal indices of refraction for quartz at this wave length are

$$n_0 = 1.5443; \qquad n_e = 1.5534$$

8. Light, polarized linearly at an angle of 30° with the optic axis of a quartz plate 0.43 mm. thick, is normally incident on a surface which is parallel to the optic axis. Assume monochromatic light of wave length 5,893 A.

a. What is the phase retardation in degrees due to the plate for the components of the electric vector parallel to and perpendicular to the optic axis?

b. What is the state of polarization of the emergent light?

9. Prove that linearly polarized light at 45° with the optic axis which is normally incident on a quarter-wave plate becomes circularly polarized upon emergence. What factors determine whether one obtains right- or left-handed circularly polarized light?

10. If two Nicols or tourmaline plates are mounted as polarizer and analyzer, prove that, when the angle between the principal sections of polarizer and analyzer is ϕ, the ratio of the intensities of the light emerging from each is given by $\cos^2 \phi$.

11. Two Nicols are used as polarizer and analyzer with their principal sections at an angle of 10° with each other. What is the relative intensity of transmitted light if the angle is changed to 75°?

12. Two light sources are observed in sequence with a polarizer and analyzer. The emergent light is found to have the same intensity for angles of 30° and 60° between the principal sections of the polarizer and analyzer, respectively.

Compute the relative intensities of the two sources.

13. Linearly polarized light of amplitude $\mathcal{E}_0$ from a Nicol prism falls normally on a quartz quarter-wave plate with the direction of $\mathcal{E}_0$ making an angle of 30° with the optic axis of the plate. The light then passes through a second Nicol oriented at 60° relative to the first Nicol.

a. What is the polarization of the light emerging from the quarter-wave plate?

b. What are the intensity and direction of polarization of the beam from the second Nicol?

CHAPTER XVI

INTERFERENCE AND DIFFRACTION

We now turn our attention to a number of optical phenomena which are entirely foreign to those which can be described with the help of geometrical optics, and these require for their explanation a more exact theory, *viz.*, the electromagnetic theory of light. In particular, experiments performed before the time of Maxwell involving interference effects (the addition of two light beams to produce darkness) made it evident that some sort of wave theory was required to explain them. We shall approach these problems from the standpoint of electromagnetic theory and at least indicate, in those cases for which the analysis becomes prohibitively involved, how one can understand these phenomena with the help of electrodynamic laws.

Fundamental in the study of wave optics is the *principle of superposition,* according to which the various wave trains which, in their totality, make up a light beam may be considered as being mutually independent, so that the behavior of the beam as a whole may be computed as the sum of the effects of the elementary waves, treating the latter as if each existed alone. This is characteristic of all wave motion and follows from the fact that the equations for wave motion are *linear* equations. The sources of electromagnetic waves of optical wave lengths are atoms, and a light beam is a complicated wave field involving the superposition of many millions of elementary wave trains emitted by the atoms in the source. The time during which an atom radiates light is of the order of 10^{-8} sec.; hence even a nearly monochromatic beam of light consists of multitudinous elementary wave trains, each about 1 meter long, of *random* polarization and phase. In particular, the randomness of the phases of these elementary waves is characteristic of any electromagnetic radiation emitted by atomic or molecular sources over which we have no individual control, and such radiation is called *incoherent* in contrast to so-called *coherent*

radiation in which the phases of the various elementary wave can be controlled to have fixed values relative to each other.

We have already mentioned the fact that one may follow the propagation of electromagnetic waves with the help of *Huygens' principle*, and we must formulate this principle (which can be derived from the fundamental electromagnetic equations) in a more precise form than we have done hitherto. In its most naïve form, Huygens' principle states that, if one knows the shape of a constant-phase surface or wave front of a wave, the position of this wave front at a later time Δt may be obtained by treating each point of the wave front as a source of secondary spherical wavelets. The wave front at time Δt is obtained by constructing spheres of radii $v\,\Delta t$ (v the velocity of the waves), using each point of the original wave surface as a center. The envelope of these wavelets gives the wave surface at time Δt. In this form there is obviously a serious defect, inasmuch as one finds a wave traveling backward as well as one traveling forward, although the former does not exist. This is illustrated in Fig. 211, in which we show a spherical wave front AA emitted by a source S, the outward traveling wave front BB at a later time Δt, and also the inward traveling wave front CC, as given by Huygens' construction.

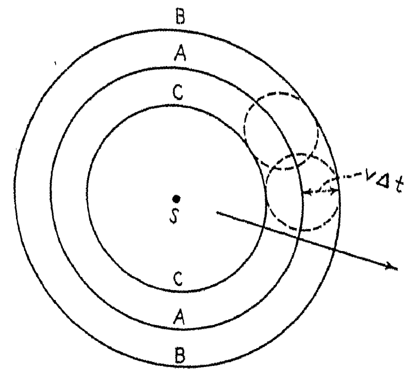

FIG. 211.

The correct formulation of Huygens' principle states that one can obtain the electromagnetic field at any point by considering each point on any *closed* surface (it may be taken as a wave front for convenience) as a source of secondary wavelets and superposing the effects of these secondary waves at the point in question. These secondary waves, however, have different amplitudes in different directions, and, when one takes the proper dependence of amplitude and phase into account, the "back" wave turns out to have zero amplitude. We shall not attempt a rigorous formulation of the principle but shall use an approximate form which is justifiable for optical wave lengths since they are always very small compared to the dimensions of the

apparatus employed in the experimental study of optical phenomena.

The fact that the resultant amplitude at a given point of an electromagnetic field is obtained by adding the amplitudes of the various waves which are passing the point in question at any instant of time gives rise to the possibility of *constructive* or of *destructive* interference, just as we have encountered in acoustics. It is usual to divide phenomena of this sort into two classes, denoted by *diffraction* or by *interference*. In diffraction effects one is concerned with the interference effects caused by a limitation of the cross section of a wave front. These effects are due to the fact that, in directions other than that of the incident wave, the mutual cancellation by destructive interference of the secondary Huygens' wavelets is not complete. On the other hand, if one causes two (or more) beams of light from *two separate portions* of the wave front to recombine, the resulting variations of intensity with position are termed *interference* effects. It must be remembered, however, that both these phenomena are fundamentally ascribable to the same process, the addition of wave amplitudes.

80. Conditions for Interference.—There are certain fundamental conditions which must be satisfied to obtain interference, some of which are inherent in the nature of light and others of which are necessary if the effects are to be observable experimentally. For simplicity let us suppose that we have two sinusoidal electromagnetic waves which are to give destructive interference at a given point through which both waves pass. If we are to have a steady interference pattern, in the sense that the intensity is to stay zero at the point in question, then it is clear that the following conditions must be satisfied:

1. The waves must have the same frequency and wave length.
2. The phase difference between the waves at the point in question must not change with time (in our case they must be 180° out of phase).
3. The amplitudes of the two waves must be equal or nearly so.
4. The waves must have the same polarization.

In the case of the interference of light waves it therefore becomes essential that the two waves which combine to give interference must come from the same source. This is due to the incoherent nature of light waves. *Light from two different*

sources can never give interference patterns. For successful observation of interference patterns produced by light, there are two more conditions which must be satisfied:

5. The difference of optical path between the two beams which combine must be very small, unless the light is monochromatic or nearly so.

6. The directions of propagation of the two interfering waves must be almost the same, *i.e.*, the wave fronts must make a very small angle with each other.

These two conditions are imposed because one deals in general with beams of light which are a mixture of many wave lengths, and it is necessary that the destructive interference at any one wave length be not masked by the partial or complete constructive interference of other wave lengths in the same beam. In Fig. 212 are shown two wave fronts traveling to the right, inclined at a slight angle to each other. Both waves have come from the same source, and let us suppose that we are dealing with a mixture of wave lengths. Suppose now that the optical path from the source to the point A is the same for both waves. At this point we have constructive interference for all wave lengths, since the phase difference between any two elementary waves of the same wave length is zero. At a point B the optical difference in path will be $\lambda/2$ for some wave length; hence we obtain destructive interference for this color. However, there will be partial constructive interference for other wave lengths at this point, but this will not amount to much for waves of wave length nearly equal to λ. On the other hand, if the path difference BB is a large integer times $\lambda/2$, one still obtains destructive interference for the color λ, but the waves of wave length almost equal to λ now completely reinforce each other and mask the effect.

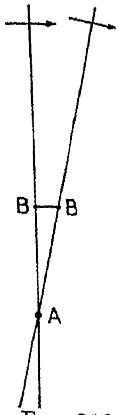

FIG. 212.

Condition 6 may be understood more clearly with the help of Fig. 213. In this figure are shown two plane waves traveling to the right with a small angle θ between them. The solid lines represent crests of the waves and the dotted lines troughs. The horizontal solid lines show the regions of complete destructive and constructive interference, *i.e.*, the widths of the so-called interference fringes. The larger the angle between the two

waves, the narrower becomes the spacing of the fringes until they cannot be separated even when magnified.

There are two general classes of devices utilized in producing interference phenomena: (1) There are those which change the direction of two parts of the same wave front so that they later recombine at a small angle. In all such devices diffraction

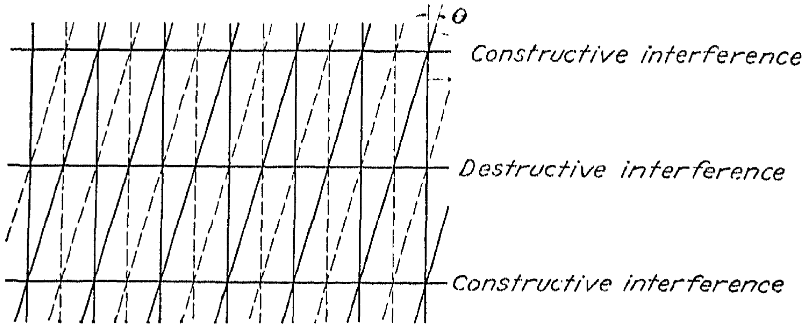

FIG. 213.

will also be present, since limited portions of the wave fronts are employed. (2) There are devices which divide the amplitude of a wave front into two parts and later reunite these two parts to produce interference. These devices may employ a large section of a wave front and minimize diffraction effects.

81. Young's Experiment.—The first experiment showing interference of light was performed by Young about 1800. The apparatus is shown schematically in Fig. 214.

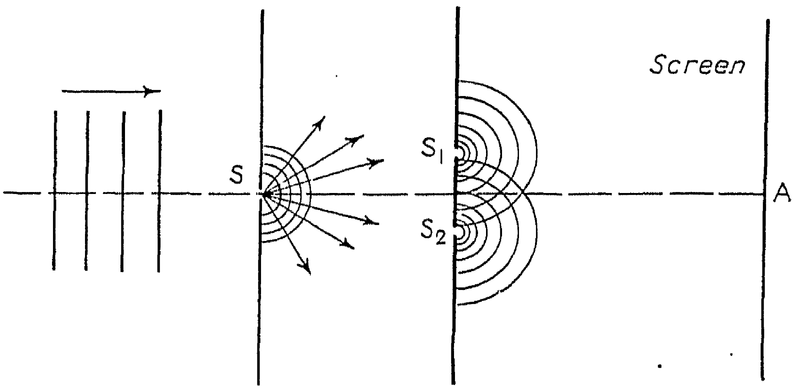

FIG. 214.

ratus is shown schematically in Fig. 214. Sunlight was allowed to pass through a pinhole S and then through two pinholes S_1 and S_2. The two spherical waves emerging from S_1 and S_2, in accordance with Huygens' principle, then interfere with each other to form an interference pattern on the screen, symmetrical

about the point A. In accordance with modern technique we shall consider the pinholes replaced by narrow parallel slits and assume that monochromatic light is employed. We now proceed to a calculation of the intensity of light at the point P of the screen as shown in Fig. 215. The slit separation d is always very small compared to D, as are the coordinates x of the points P at which the pattern is observed. Thus we have

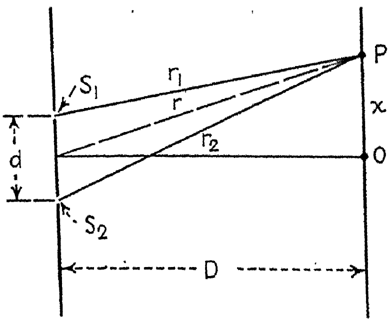

FIG. 215.

$$d << D \qquad \text{and} \qquad x << D.$$

If the slits S_1 and S_2 are equidistant from the source slit S, the electric vectors at these slits will be in phase and of equal amplitude. Either may be written as

$$A \sin 2\pi\nu t$$

The electric vector at P will be the sum of the amplitudes of the waves coming from S_1 and from S_2 according to Huygens' principle. Thus we have

$$\mathcal{E}_P = \mathcal{E}_1 \sin 2\pi\left(\nu t - \frac{r_1}{\lambda}\right) + \mathcal{E}_2 \sin 2\pi\left(\nu t - \frac{r_2}{\lambda}\right) \tag{1}$$

which may be written in the form

$$\mathcal{E}_P = \mathcal{E}_0 \sin (2\pi\nu t - \phi) \tag{2}$$

with

$$\mathcal{E}_0^2 = \mathcal{E}_1^2 + \mathcal{E}_2^2 + 2\mathcal{E}_1\mathcal{E}_2 \cos \frac{2\pi(r_2 - r_1)}{\lambda} \tag{3}$$

as one may show by vector addition of the two sine waves of Eq. (1). The phase ϕ is of no interest in the calculation of relative intensities. Since r_1 and r_2 are practically equal and are very large compared to all values of x in which we are interested, we may place $\mathcal{E}_1 = \mathcal{E}_2 = \mathcal{E}$, thus neglecting the variation of amplitude of the Huygens' wavelets with direction and distance, a procedure which is justifiable in this case. Equation (3) then becomes

$$\mathcal{E}_0^2 = 2\mathcal{E}^2\left[1 + \cos \frac{2\pi}{\lambda}(r_2 - r_1)\right] = 4\mathcal{E}^2 \cos^2 \frac{\pi(r_2 - r_1)}{\lambda} \tag{4}$$

The angle $(2\pi/\lambda)(r_2 - r_1)$ is just the phase difference of the

two waves arriving at P from S_2 and S_1, respectively. Since the intensity is proportional to $\mathcal{E}_0^2$, we may rewrite Eq. (4) in the following form, calling I_1 the intensity due to one slit,

$$\frac{I}{I_1} = 4 \cos^2 \frac{\pi(r_2 - r_1)}{\lambda} \tag{5}$$

There remains the task of expressing $(r_2 - r_1)$ in terms of the distances shown in Fig. 215. We have

$$r_2^2 = D^2 + \left(x + \frac{d}{2}\right)^2$$
$$r_1^2 = D^2 + \left(x - \frac{d}{2}\right)^2$$

so that

$$r_2^2 - r_1^2 \cong 2r(r_2 - r_1) = 2xd \tag{6}$$

where we have set $r_1 + r_2 = 2r$. Since $x << D$, we may replace r by D without appreciable error and have finally

$$r_2 - r_1 = \frac{d}{D} \cdot x \tag{7}$$

so that Eq. (5) becomes

$$\frac{I}{I_1} = 4 \cos^2 \left(\frac{\pi d}{\lambda D} \cdot x\right) \tag{8}$$

which gives the intensity distribution along the x-axis on the screen. The intensity is maximum for

$$\frac{\pi d}{\lambda D} x = k\pi$$

or at the points

$$x_k = \frac{k\lambda D}{d} \qquad (k = 0, 1, 2, \cdots) \tag{9}$$

so that the maxima are uniformly spaced.

The minima of the intensity are of intensity zero and lie at the positions given by

$$\frac{\pi d}{\lambda D} x = (k + \tfrac{1}{2})\pi \qquad (k = 0, 1, 2, 3, \cdots)$$

or

$$x_k = (k + \tfrac{1}{2})\frac{\lambda D}{d} \tag{10}$$

The integer k is called the order of the interference. The distance between two neighboring maxima (or minima) gives the spacing S of the fringes and is equal to

$$S = \frac{\lambda D}{d} \tag{11}$$

and this affords a direct method of determining the wave length of light.

The maxima and minima may be located by an elementary argument. Maxima will occur whenever the difference of path $(r_2 - r_1)$ is equal to a whole number of wave lengths, giving us constructive interference. Thus we must have $(r_2 - r_1) = k\lambda$, or, using Eq. (7),

$$x_k = \frac{k\lambda D}{d}$$

in agreement with Eq. (9). The location of the minima proceeds in exactly the same manner, setting $(r_2 - r_1)$ equal to an odd number of half wave lengths so as to produce destructive interference.

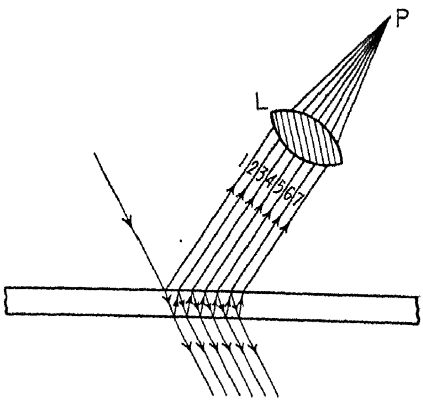

FIG. 216.

82. Interference in Thin Films; Newton's Rings.—The brilliant colors of a thin film of oil floating on water or of a thin soap film are due to interference effects of the type where the amplitude of the incident wave is divided by reflection and refraction at the boundaries of the film. Suppose a plane wave in air is incident on the film. Part of the wave is reflected at the first surface and part transmitted. This second wave is partially reflected (internally) at the second interface and partially transmitted through the film. When the internally reflected wave impinges on the first surface, part of it is transmitted and part internally reflected, and so on. The phase differences between the first reflected wave and those emerging from the top surface after successive internal reflections give rise to interference patterns. The multiple reflections between the boundaries of a thin film with parallel faces is illustrated in

Fig. 216, the various reflected rays being focused at P by the lens L.

To find the phase differences between these reflected waves, we must compute the path difference for a pair of successive waves, such as 1 and 2 in Fig. 216. In so doing we must remember that there is a phase difference of 180° between the incident and reflected electric vector produced when a wave is reflected at the surface of an optically denser medium and that no phase change occurs when reflection takes place at the surface of an optically rarer medium [see Eq. (25), Chap. XIII]. Thus we can understand why a very thin soap film (of thickness small compared to the wave length of light) appears black by reflected light.

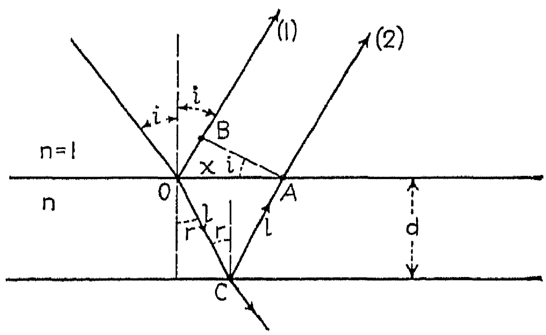

FIG. 217.

In Fig. 217 let d be the thickness of the film of refractive index n, and let i and r be the angles of incidence and refraction, as shown. We are to calculate the phase difference between the corresponding points A and B on the reflected rays, where AB is drawn perpendicular to the rays (1) and (2). The distance traveled by the internally reflected ray OCA is evidently

$$2l = \frac{2d}{\cos r}$$

The time of traversal of the path OCA is hence

$$t_2 = \frac{2l}{c/n} = \frac{2nd}{c \cos r} \tag{12}$$

The time of traversal of the distance $OB = x \sin i = 2l \sin r \sin i$ is similarly

$$t_1 = \frac{2l \sin r \sin i}{c} \tag{13}$$

The difference of time of traversal gives rise to a phase difference equal to $2\pi\nu(t_2 - t_1)$, from which we must subtract π since the electric vector undergoes this change of phase at reflection of ray (1), whereas there is no such change at point C for internal reflection. The difference of time is thus

$$t_2 - t_1 = \frac{2l}{c}(n - \sin i \sin r)$$

or, since $\sin i = n \sin r$,

$$t_2 - t_1 = \frac{2nl}{c}(1 - \sin^2 r) = \frac{2nd}{c}\cos r \tag{14}$$

and the phase difference becomes

$$\Delta\phi = \frac{4\pi\nu n d}{c}\cos r - \pi$$

or, since $\nu/c = 1/\lambda$,

$$\Delta\phi = \frac{4\pi n d}{\lambda}\cos r - \pi \tag{15}$$

The necessary condition for interference maxima is that $\Delta\phi = 2\pi k$, where k is an integer, so that for maxima

$$\frac{4\pi n d}{\lambda}\cos r - \pi = 2k\pi$$

or

$$2nd\cos r = (k + \tfrac{1}{2})\lambda \qquad \text{(maxima)} \tag{16}$$

whereas for minima

$$2nd\cos r = k\lambda \qquad \text{(minima)} \tag{17}$$

where, in Eqs. (16) and (17), k is any integer or zero.

If Eq. (17) is satisfied, so that rays 1 and 2 are out of phase, it is easy to see that the other reflected rays 3, 4, 5, etc., of Fig. 216 emerge in phase with 2. This follows from the fact that the phase difference between any succeeding rays, such as 2 and 3, is evidently given by $(2\pi/\lambda)(2nd \cos r)$, and, if Eq. (17) is satisfied, this is $2\pi k$. On the other hand, when the condition of constructive interference, as given by Eq. (16), is attained for rays 1 and 2, we see that rays 2, 4, 6, etc., will be in phase with 1, whereas rays 3, 5, 7, etc., will be half a wave length out of phase with 1. Since the amplitude drops off strongly on successive reflections, there will still be a maximum intensity under these conditions. For the minima of intensity, ray 2 is considerably weaker than ray 1, so that it alone cannot completely annul it. However, one can show that the sum of the amplitudes of all the successive waves 2, 3, 4, etc., is just equal to the amplitude of wave 1, yielding complete darkness at the minima.

Equations (16) and (17) show that for normal incidence ($\cos r = 1$) strong reflection occurs when the film thickness is an odd multiple of a quarter wave length in the film, whereas no reflection occurs if the thickness is an even number of quarter wave lengths. This is the principle underlying the behavior of so-called "invisible" glass made by evaporating a thin transparent film on its surface.

If the convex surface of a plano-convex lens is placed in contact with a plane glass plate, a thin film of air of varying thickness will be formed between the surfaces. The loci of points of equal thickness will be circles concentric with the point of contact. Such an air film shows circular interference bands, known as *Newton's rings*. When viewed by reflected light, the center of the pattern is black and, when viewed by transmitted light, it is white.

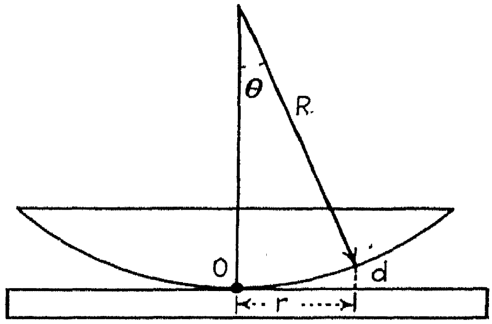

FIG. 218.

To obtain a relation between the radii of the interference rings, the wave length, and the curvature of the lens surface, we have from Fig. 218

$$d = R(1 - \cos\theta) = 2R\sin^2\frac{\theta}{2}$$

For small angles θ we may set $\sin^2(\theta/2)$ equal to $\theta^2/4$ and place $\theta = r/R$ so that d is very nearly given by

$$d = \frac{\quad}{2R} \tag{18}$$

Hence we shall observe a dark ring by reflected light if, according to Eq. (17) ($n = 1$ for the air film),

$$2d\cos r = k\lambda$$

so that, using Eq. (18) and placing $\cos r = 1$, since the angle r as measured with normal to the film surface is very small,

$$\frac{r^2}{R} = k\lambda$$

or

$$r_k^2 = kR\lambda \qquad (k = 0, 1, 2, \cdots) \tag{19}$$

giving the radii of the dark rings. Similarly, the radii of the bright fringes are given, according to Eq. (16), by

$$r_k^2 = (k + \tfrac{1}{2})R\lambda \qquad (k = 0, 1, 2, \cdots) \tag{20}$$

83. Interferometers.—The *Michelson interferometer* is an instrument which can be used to measure exceedingly small distances in terms of the wave length of light. The essential parts are shown in Fig. 219. Light from a source S is collimated by lens L and falls on a plate P_1 which is inclined at an angle of 45° with the beam and is lightly coated with silver on its back surface, so that approximately half the light is transmitted to

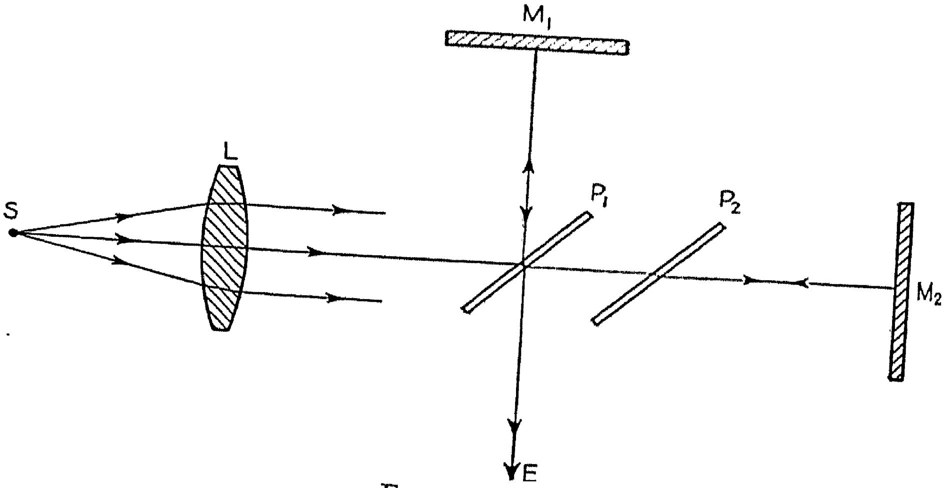

FIG. 219.

mirror M_2 and the other half is reflected to mirror M_1. The plate P_2 is identical with and parallel to plate P_1, except that it is not silvered. It is used so that the optical paths $P_1M_1P_1$ and $P_1M_2P_1$ contain the same thickness of glass. This is important whenever light of many wave lengths is used because of the dispersion of glass.

The interference pattern is observed at E with the help of a telescope. One sees the surface of the mirror M_1 through the half-silvered plate P_1 and the surface of the mirror M_2 reflected in P_1. If the distances from P_1 to the two mirrors are exactly equal and if the mirrors M_1 and M_2 are exactly at right angles to each other and at 45° with P_1, the image of M_2 coincides with the surface of M_1. If the adjustment is not exact, then in effect a thin air film exists between the surface of M_1 and the image of M_2 and causes the interference pattern observed. As the

mirror M_2, let us say, is moved, the system of fringes is displaced, and a displacement of the mirror of one-half wave length causes each fringe to move to the position formerly occupied by an adjacent fringe. Thus, by counting fringes, extremely small distances may be measured.

The Fabry-Perot interferometer utilizes the fringes produced in the light transmitted by an air film between two lightly silvered surfaces of plane parallel plates P_1 and P_2 of Fig. 220. The separation d between the reflecting surfaces is of the order of one centimeter, and the observations are made near normal incidence. To observe the fringes, monochromatic light from an extended source, of which S is one point, is made parallel by the lens L_1, and the transmitted light is brought together to produce inter-

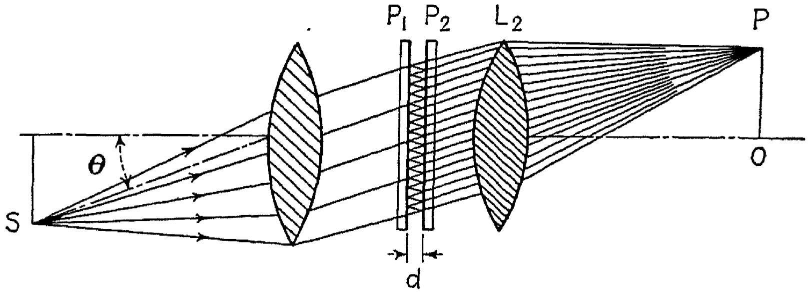

Fig. 220.

ference by the lens L_2. In Fig. 220 a ray from S is incident at an angle θ, producing a series of parallel rays at the same angle, which are focused at P by the lens L_2. The condition for constructive interference will be the same for all points on a circle of radius OP, so that circular fringes will be observed. In the actual instrument one of the plates is fixed and the other may be moved toward or away from it to vary the distance d.

84. Fresnel and Fraunhofer Diffraction; Fresnel Zones.—When the cross section of a beam of light is limited by allowing the light to pass through an opaque screen containing one or more apertures, the distribution of intensity in the transmitted beam as observed on another screen or with the help of a telescope is called a diffraction pattern. If the diffracting screen (or obstacle) is placed between source and observing screen and *no* lenses or mirrors are employed, the resulting phenomenon is called *Fresnel* diffraction. In general, both source and observing screen are at finite distances from the diffracting screen. If, on the other hand, one employs a plane wave of incident light,

either from a distant source or collimated with the help of a lens, and the diffracted waves are observed with the help of a telescope focused on infinity or observed on a very distant screen, the resulting pattern is known as a *Fraunhofer diffraction pattern.* Fundamentally, both types of diffraction are only different aspects of the same basic phenomenon and are explicable in terms of Huygens' principle. Our first task, then, is to examine this principle more closely than we have up to this point and to see how it explains the rectilinear propagation of light for unobstructed waves.

Consider a spherical wave diverging from a source O, and suppose we wish to compute the amplitude of this wave at a point P, which lies at a distance R from the source O, utilizing

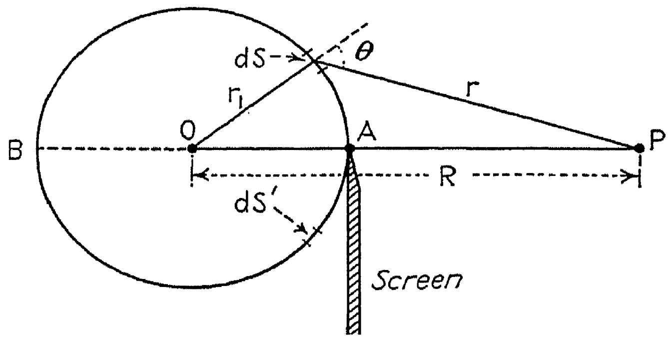

FIG. 221.

Huygens' principle. First we construct a spherical surface of radius $r_1 < R$ with its center at O. This is a wave front, and we must consider each element of area dS on this surface as the source of secondary waves which in their totality combine at P to give the resultant wave motion at P (Fig. 221). The relative phases of the secondary waves arriving at P may be readily obtained by noting that it takes a time r/c for a disturbance at dS to reach P. Thus the relative phases of these waves are given by $2\pi r/\lambda$. It is not evident what the relative amplitudes of these waves will be. We would expect them to be proportional to dS, the area of the elementary source on the wave front, inversely proportional to r, and this is true. In addition, however, it turns out that they depend on the angle θ in Fig. 221 in the form $(1 + \cos\theta)$, where θ is the angle between the normal to the spherical surface and r, so that $\cos\theta$ varies from $+1$ at A to -1 at B. This so-called *obliquity* factor eliminates the "back" wave in the elementary Huygens' construction.

As we shall see presently, for an unobstructed wave the secondary waves from all elements of area dS mutually destroy each other by interference at P, *except* for those originating in a very small region around the point A; hence to all intents and purposes the effect is the same as if the light traveled in a straight line from O to P.

If now some sort of obstacle, such as the opaque screen indicated in Fig. 221, is inserted between O and P, we no longer have the possibility of the mutual cancellation of the secondary waves from dS and dS' (let us say), since the screen prevents the waves from dS' from reaching P. This gives rise, then, to diffraction, and one may obtain even larger intensities at P

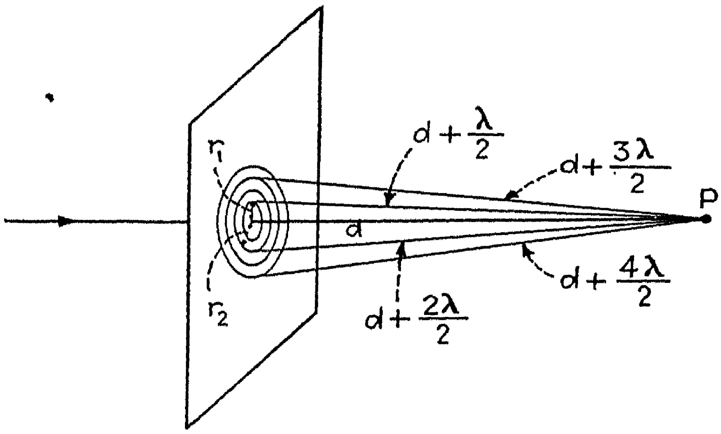

FIG. 222.

than without the screen. *Fresnel* has given an ingenious method of computing approximately the contributions from the various secondary waves which enables one to obtain the essential results without complicated integrations. Let us consider this method for the case of a plane wave to compute the effect produced at a point P ahead of the wave (Fig. 222). In a plane wave front we describe a series of circles about O as a center (Fig. 222) of radii r_1, r_2, r_3, etc., such that the distances from P to these various circles increase by $\lambda/2$ as we go from one circle to its neighbor. The line $PO = d$ is perpendicular to the plane. We have thus divided the wave front into zones called *Fresnel zones* or *half-period elements*, and a similar construction may be readily carried out for the case of a spherical wave, as in Fig. 221. Now the phase difference between the waves arriving at P from O and from the edge of the first zone (at r_1) is just

$$2\pi\left[\frac{d + (\lambda/2)}{\lambda} - \frac{d}{\lambda}\right] = \pi$$

so that all the waves arising from points lying within this first zone give contributions, at a given instant of time, of the same algebraic sign (let us say positive). Similarly, the phases of the waves coming from the second zone (relative to that coming from O) lie between π and 2π, and these waves give negative contributions, those from the third zone positive contributions, and so on. The resultant amplitude A at P will then be the sum of the contributions from the various Fresnel zones, a sum of terms of the form

$$A = a_1 - a_2 + a_3 - a_4 + \cdot \cdot \cdot + a_n \tag{21}$$

where successive terms alternate in sign and have magnitudes which decrease very slowly, as we shall see, as we proceed from one term to the next. The magnitudes a_n vary because of three effects: (1) the areas of the zones change slightly from zone to zone, (2) the distance from the zones to P increases slightly with increasing zone number, and (3) the angle θ referred to in the obliquity effect increases slowly with increasing zone number. The net effect is that there is a slow decrease of magnitude of a_n with n. This is also true for the spherical wave case.

Let us compute the areas of these zones for the case under discussion. From Fig. 222 we see that, for the nth circle, we have

$$d^2 + r_n^2 = \left(d + \frac{n\lambda}{2}\right)^2 = d^2 + n\lambda d + \frac{n^2\lambda^2}{4}$$

or

$$r_n^2 = nd\lambda + \frac{n^2\lambda^2}{4} \tag{22}$$

and for the $(n - 1)$st circle

$$r_{n-1}^2 = (n - 1)d\lambda + (n - 1)^2 \frac{\lambda^2}{4} \tag{23}$$

The area of the nth zone is accordingly

$$S_n = \pi r_n^2 - \pi r_{n-1}^2 = \pi d\lambda + \frac{\pi}{2}(n - \tfrac{1}{2})\lambda^2$$

or

$$S_n = \pi d\lambda\left(1 + \frac{n - \frac{1}{2}}{2}\frac{\lambda}{d}\right) \tag{24}$$

Now, in general, $\lambda << d$, so that the second term inside the parentheses of Eq. (24) is negligible, and we have very nearly

$$S_n = \pi d\lambda \tag{25}$$

independent of the zone number.

Returning to the question of the amplitude A as given by Eq. (21), we shall show that it is very nearly given by

$$A = \tfrac{1}{2}(a_1 + a_n) \tag{26}$$

i.e., that the sum of an alternating series is approximately half the sum of the first and last terms if the magnitudes of successive terms are almost equal. To see this, let us rewrite Eq. (21) in the form

$$A = \frac{a_1}{2} + \left(\frac{a_1}{2} - a_2 + \frac{a_3}{2}\right) + \left(\frac{a_3}{2} - a_4 + \frac{a_5}{2}\right) + \cdots + \left(\frac{a_{n-2}}{2} - a_{n-1} + \frac{a_n}{2}\right) + \frac{a_n}{2}$$

and, since the amplitude from any zone is nearly equal to the average of those of the preceding and following zones, we may write

$$a_2 = \frac{a_1 + a_3}{2}$$

$$a_4 = \frac{a_3 + a_5}{2}$$

etc.,

so that all the terms in the parentheses vanish, and we are left with the result expressed by Eq. (26).

If we are concerned with a problem in which a very large number of zones contribute, the effect of the last zone is in general negligible, and we have

$$A = \frac{a_1}{2} \tag{27}$$

showing that the amplitude at P is essentially one-half the contribution from the first zone. Thus, for example, in the case of the unobstructed spherical wave of Fig. 221, the amplitude at P may be thought of as coming practically from the point A of that figure, so that the concept of a ray is justified from the standpoint of wave theory.

85. Application of Fresnel Zones to Fresnel Diffraction.—The general method of calculation of the diffraction pattern due to the interposition of a plane opaque screen, in which there are one or more apertures between the source and the observing screen, is to assume that the amplitude of the light wave at all points of the apertures is the same as if the diffracting screen were absent and then to sum up the contributions at the observing screen of the various Huygens' secondary waves emitted from the points within the apertures. For the Fresnel case this is a complicated task, and we shall not attempt quantitative solutions, but shall examine the qualitative nature of the phenomena with the help of Fresnel zones. Suppose, for simplicity, that there is a single aperture in the diffracting screen in the form of a rectangular slit. In this screen we imagine the Fresnel zones constructed and can then examine the diffraction pattern, the nature of which depends on which zones are uncovered and can transmit waves and on which zones are obscured by the screen.

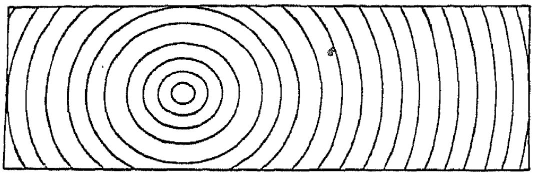

FIG. 223.

Suppose first that the center of the zone system lies well within the aperture, corresponding to an observation point P which should be fully illuminated according to geometrical optics. The central zone is fully uncovered as well as a number of others, but, as we proceed to larger zones, they become partially covered and finally completely covered. This is indicated in Fig. 223. The amplitude A at the observation point is the sum of contributions from the various exposed zones, and, while the areas decrease somewhat more rapidly when we reach the partly covered zones than for an unobstructed wave, we still may use the result that A is given by half the sum of the first and last terms. Since the last zone is almost entirely covered, we have left just half the contribution from the first zone, so that we obtain the same intensity as if the screen were absent.

As a second case, suppose the observation point P is near the edge of the geometrical shadow, making the center of the zone system near the edge of the aperture. The first zone may be partly obscured so that the intensity is less than without the screen. If the first zone is completely uncovered but the next

ones partially obscured, the contributions a_2, a_3, etc., may vary so rapidly that it would be incorrect to take half the sum of the first and last zone contributions. In such a case one might well obtain an amplitude A greater than $a_1/2$, so that the intensity would be *greater* than in the absence of the screen. As one moves past the edge, successive zones become covered, and there is periodic variation of intensity; these are the diffraction fringes.

Finally, if we move the point P well into the geometrical shadow, the first few zones are obscured. A certain zone is partially uncovered and succeeding zones become more uncovered to a considerable extent. The larger zones again become more and more obscured, and, in the sum of the contributions $a_1 \cdots a_n$, the first and last terms are zero, so there is no intensity

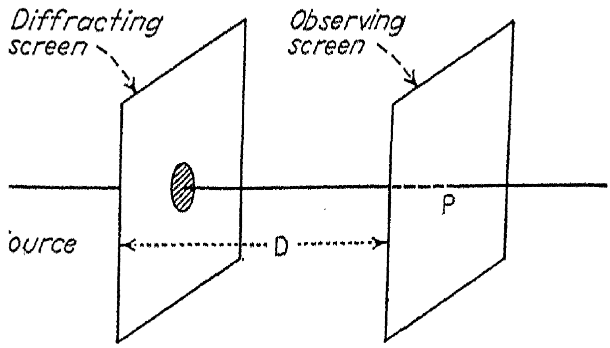

FIG. 224.

at this point. This is well within the geometrical shadow, and, speaking physically, the waves from the partially uncovered zones mutually cancel by interference.

Having obtained a general picture, let us examine a few special cases more closely. Consider first the diffraction pattern formed by a circular opening in the diffracting screen as shown in Fig. 224. Suppose that the size of the circular opening is such that only the first zone (the point P being on the axis) is uncovered. Then the amplitude at P would be a_1, or *twice* the amplitude in the absence of the screen. This means that the intensity is four times as great. If we imagine the opening now increased until two zones are uncovered, we have an amplitude given by $a_1 - a_2$, and this is practically zero since a_1 and a_2 are almost equal to each other. Thus it becomes clear that the intensity at P is a maximum when an *odd* number of zones is uncovered and a minimum when an *even* number of zones is uncovered. The same effect may be obtained by moving the observation screen

with a *fixed* aperture. We have seen [Eq. (22)] that the radius of the first zone is given by $\sqrt{\lambda D}$, and, as D is varied, one can obscure or uncover a larger number of zones, leading to alternating intensities along the axis. Similarly, one can follow the alternations in intensity as the point P moves laterally into the geometrical shadow. One can also discuss the diffraction pattern caused by a circular obstacle by exactly the same scheme. One obtains the surprising result that, if the obstacle obscures only a few Fresnel zones, a bright spot should appear at the center of the shadow. This has been observed experimentally and provides a most convincing argument in favor of a wave theory of light. Diffraction by a straightedge may be examined with the help of an appropriate Fresnel zone construction, differing only in detail from the examples given above. Details are left to the problems.

86. Fraunhofer Diffraction.—We shall now examine the Fraunhofer diffraction pattern produced by a single rectangular slit. For the sake of simplicity let the diffracting screen containing the slit be perpendicular to the incident plane wave of monochromatic light. The pattern is observed on a screen very far from the slit (at infinity) or with the help of a telescope focused on infinity, so that the intensity at any point of the pattern is due to the superposition of all the diffracted rays leaving various points of the aperture in a *given direction*. Let us choose a coordinate system, as shown in Fig. 225, with the origin at the center of the rectangular slit of width a and length b. We start with the observation screen at a finite distance z_0 from the slit and consider the wave arriving at P (coordinates x_0, y_0, z_0) as the superposition of elementary waves coming from the various infinitesimal elements of area dS of the aperture. Since we are only interested in the limiting case of infinite distance from slit to P, we may take the relative amplitudes of the various waves arriving from elementary areas dS (of equal size) as equal. In so doing, we neglect the slight variation with angle (the obliquity effect); consequently our results will be valid for small angles, *i.e.*, near the center of the pattern O'. The variations of intensity in the pattern are thus due practically only to the relative differences of phase among the various waves. The contribution du to one of the components of the electric field, let us say, at P from the wave coming from dS is

$$du = A\,dS \sin 2\pi\left(\nu t - \frac{r}{\lambda}\right) \tag{28}$$

corresponding to an oscillation

$$A'\,dS \sin 2\pi\nu t$$

at dS, and the sum of all the contributions from the various elementary areas dS of the slit is clearly

$$u = \int\int A \sin 2\pi\left(\nu t - \frac{r}{\lambda}\right) dS \tag{29}$$

where the integration is carried out over the whole slit area. In order to carry out the integration, we must express r, the

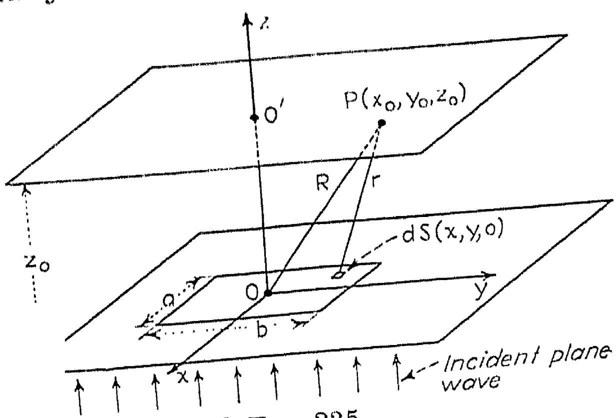

FIG. 225.

distance from dS to P, in terms of the coordinates of dS and of P. We have

$$r^2 = (x - x_0)^2 + (y - y_0)^2 + z_0^2$$

or, expanding and placing $x_0^2 + y_0^2 + z_0^2 = R^2$ (see Fig. 225),

$$r^2 = R^2 - 2xx_0 - 2yy_0 + x^2 + y^2 \tag{30}$$

Now x_0/R is the cosine of the angle between R and the x-axis, which we denote by l, and $y_0/R = m$ is the cosine of the angle between R and the y-axis. Using this notation, Eq. (30) becomes

$$r^2 = R^2\left(1 - \frac{2lx + 2my}{R} + \frac{x^2 + y^2}{R^2}\right) \tag{31}$$

Now x/R is very small compared to unity if R becomes very large (the Fraunhofer case), so that we may neglect terms of the order

x^2/R^2, y^2/R^2, and higher powers. Therefore we drop the last term in the parentheses of Eq. (31), take the square root, and expand according to the binomial theorem. The result is

$$r = R - (lx + my) \tag{32}$$

Setting $dS = dx\,dy$ and inserting Eq. (32) into Eq. (29), there follows

$$u = A\int_{-\frac{a}{2}}^{+\frac{a}{2}}\int_{-\frac{b}{2}}^{+\frac{b}{2}} \sin 2\pi\left(\nu t - \frac{R}{\lambda} - \frac{lx + my}{\lambda}\right) dx\,dy \tag{33}$$

If we denote $2\pi\left(\nu t - \dfrac{R}{\lambda}\right)$ by ϕ, the integrand may be written in the form

$$\sin\left[\left(\phi - \frac{2\pi lx}{\lambda}\right) - \frac{2\pi my}{\lambda}\right] = \sin\left(\phi - \frac{2\pi lx}{\lambda}\right)\cos\frac{2\pi my}{\lambda} - \cos\left(\phi - \frac{2\pi lx}{\lambda}\right)\sin\frac{2\pi my}{\lambda} \tag{34}$$

When integrated with respect to y, the second term on the right-hand side of Eq. (34) gives zero, since

$$\left[\cos\frac{2\pi my}{\lambda}\right]_{-\frac{b}{2}}^{+\frac{b}{2}} = 0$$

The first term on the right-hand side of Eq. (34) may itself be rewritten as

$$\sin\left(\phi - \frac{2\pi lx}{\lambda}\right)\cos\frac{2\pi my}{\lambda} = \sin\phi\cos\frac{2\pi lx}{\lambda}\cos\frac{2\pi my}{\lambda} - \cos\phi\sin\frac{2\pi lx}{\lambda}\cos\frac{2\pi my}{\lambda} \tag{35}$$

and the second term becomes zero when integrated with respect to x, since

$$\left[\cos\frac{2\pi lx}{\lambda}\right]_{-\frac{a}{2}}^{+\frac{a}{2}} = 0$$

Thus we are left with the first term of Eq. (35) which, when reinserted into Eq. (33), gives us

$$u = A \sin 2\pi\left(\nu t - \frac{R}{\lambda}\right)\int_{-\frac{a}{2}}^{+\frac{a}{2}} \cos \frac{2\pi lx}{\lambda}\, dx \int_{-\frac{b}{2}}^{+\frac{b}{2}} \cos \frac{2\pi my}{\lambda}\, dy \quad (36)$$

the product of two identical integrals. The integral with respect to x becomes

$$\frac{\lambda}{2\pi l}\left[\sin \frac{2\pi lx}{\lambda}\right]_{-\frac{a}{2}}^{+\frac{a}{2}} = \frac{\lambda}{\pi l} \sin \frac{\pi la}{\lambda}$$

and similarly the other one is

$$\frac{\lambda}{2\pi m}\left[\sin \frac{2\pi my}{\lambda}\right]_{-\frac{b}{2}}^{+\frac{b}{2}} = \frac{\lambda}{\pi m} \sin \frac{\pi mb}{\lambda}$$

so that Eq. (36) yields

$$u = A \cdot ab \cdot \sin 2\pi\left(\nu t - \frac{R}{\lambda}\right) \frac{\sin (\pi la/\lambda)}{\pi la/\lambda} \cdot \frac{\sin (\pi mb/\lambda)}{\pi mb/\lambda} \quad (37)$$

and, since the intensity of light is proportional to the square of u, we may write

$$\frac{I}{I_0} = \frac{\sin^2 \alpha}{\alpha^2} \cdot \frac{\sin^2 \beta}{\beta^2} \quad (38)$$

where we have set

$$\alpha = \frac{\pi la}{\lambda}; \qquad \beta = \frac{\pi mb}{\lambda} \quad (38a)$$

and I_0 is the intensity for $\alpha = \beta = 0$, *i.e.*, at the point O' in Fig. 225. Equation (38) gives the variation of intensity with position (given by the direction cosines l and m) in the pattern. Since the variation is identical in the x-z plane (transverse to the length of the slit) with that in the y-z plane (along the slit length), it will be sufficient to examine the variation of intensity in the x-z plane. For all points P in this plane, $m = 0$ since the angle between R (of Fig. 225) and the y-axis is 90°. Thus $\beta = 0$ and

$$\frac{\sin^2 \beta}{\beta^2} = 1$$

For this case, Eq. (38) becomes

$$\frac{I}{I_0} = \frac{\sin^2 \alpha}{\alpha^2} \quad (39)$$

and this is shown plotted in Fig. 226. The maximum intensity occurs at the center, falls off to zero for $\alpha = \pi$, 2π, 3π, etc., with secondary maxima approximately halfway between. If we take the positions of the maxima at the points

$$\alpha = \frac{\pi a l}{\lambda} = \frac{3\pi}{2}, \frac{5\pi}{2}, \text{ etc.,}$$

the relative intensities at these points are

$$\left(\frac{2}{3\pi}\right)^2; \left(\frac{2}{5\pi}\right)^2; \left(\frac{2}{7\pi}\right)^2; \text{ etc.,} \qquad \text{or} \qquad 0.045; 0.016; \text{ etc.}$$

Thus we see that the intensities of the secondary maxima fall off very rapidly as one proceeds away from the central maximum,

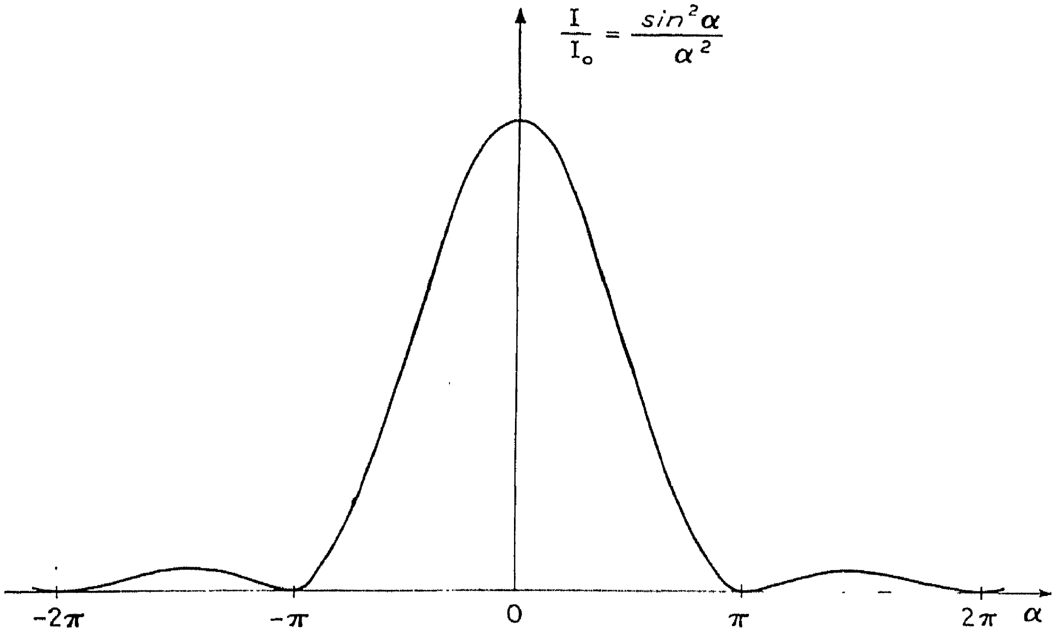

FIG. 226.

so that practically all the light is concentrated in the central diffraction band. The half angle θ subtended at the slit by this band is given by

$$\sin\theta = l = \lambda$$

or, for small angles,

$$\theta = \frac{\lambda}{a} \tag{40}$$

This is shown in Fig. 227, in which we may imagine the slit dimension b to be so large compared to a that we have essentially a one-dimensional pattern. Note that the pattern becomes more extended the longer the wave length, or the narrower the slit dimensions. The general expression for the location of the minima is given by

$$\frac{\pi a l}{\lambda} = k\pi \qquad (k = 1, 2, 3, \qquad)$$

or for the angles

$$\sin \theta = l = k\frac{\lambda}{a} \qquad (k = 1, 2, 3, \qquad) \tag{40a}$$

In considering Young's experiment on the interference of light by two narrow slits close together, the assumption was

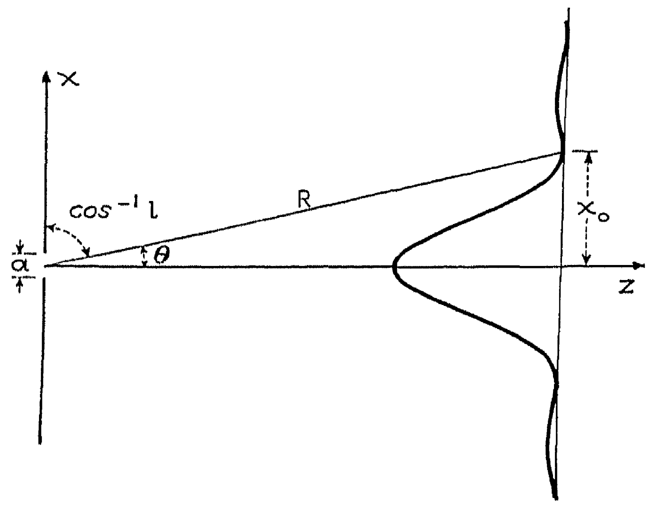

FIG. 227.

tacitly made that the slits were narrow enough and close enough together to cause considerable overlapping of the central maxima of the diffraction patterns of each slit. We shall now examine the problem of the Fraunhofer diffraction pattern due to *two slits*, each of width a and separation d between centers. We shall, for simplicity, consider the slit lengths large compared to their widths, so that we may be concerned only with the one-dimensional pattern, *i.e.*, with the variation of intensity in the x-direction on the observation screen. Even for slits not long compared to their width, the pattern is altered from the single slit pattern only in this dimension, so that we lose no generality

by taking $m = 0$ ($y_0 = 0$) for the points P at which we compute the intensity.

The method of calculation is identical with that employed for a single slit, except that we must now carry out the integration over two slits instead of one. Thus we may take over Eq. (33) as it stands, setting $m = 0$, and have

$$u = Ab\left[\int_{-\frac{a}{2}}^{+\frac{a}{2}} \sin 2\pi\left(\nu t - \frac{R}{\lambda} - \frac{lx}{\lambda}\right) dx + \int_{d-\frac{a}{2}}^{d+\frac{a}{2}} \sin 2\pi\left(\nu t - \frac{R}{\lambda} - \frac{lx}{\lambda}\right) dx\right] \quad (41)$$

since, with an origin at the center of the first slit, the second extends from $\left(d - \frac{a}{2}\right)$ to $\left(d + \frac{a}{2}\right)$. If in the second integral we set $x' = x + d$, it becomes

$$\int_{-\frac{a}{2}}^{+\frac{a}{2}} \sin 2\pi\left(\nu t - \frac{R}{\lambda} - \frac{ld}{\lambda} - \frac{lx'}{\lambda}\right) dx'$$

Proceeding as before, each term in Eq. (41) yields a term of the form of Eq. (36) with $m = 0$, and we obtain in place of Eq. (36),

$$u = Ab\left(\int_{-\frac{a}{2}}^{+\frac{a}{2}} \cos \frac{2\pi lx}{\lambda}\, dx\right)\left[\sin 2\pi\left(\nu t - \frac{R}{\lambda}\right) + \sin 2\pi\left(\nu t - \frac{R}{\lambda} - \frac{ld}{\lambda}\right)\right] \quad (42)$$

The term in the square brackets may be written in the form

$$B \sin (2\pi\nu t - \delta)$$

with

$$B^2 = 2\left(1 + \cos \frac{2\pi ld}{\lambda}\right) = 4 \cos^2 \frac{\pi ld}{\lambda} \quad (43)$$

The phase δ does not interest us since it does not contribute to the relative intensity. Note that Eq. (43) corresponds exactly to Eq. (8) for the interference pattern relative intensity in Young's experiment, since $l = x/D$. The remaining terms

in Eq. (42) yield exactly the relative intensity in the diffraction pattern of a single slit as expressed in Eq. (39). Thus we may write for the relative intensity in the double slit pattern

$$\frac{I}{I_0} = 4\frac{\sin^2\alpha}{\alpha^2}\cos^2\gamma \tag{44}$$

with $\alpha = \pi la/\lambda$ as before, and $\gamma = \pi ld/\lambda$.

Thus the intensity pattern is given by the product of two factors, one the diffraction pattern of a single slit and the other

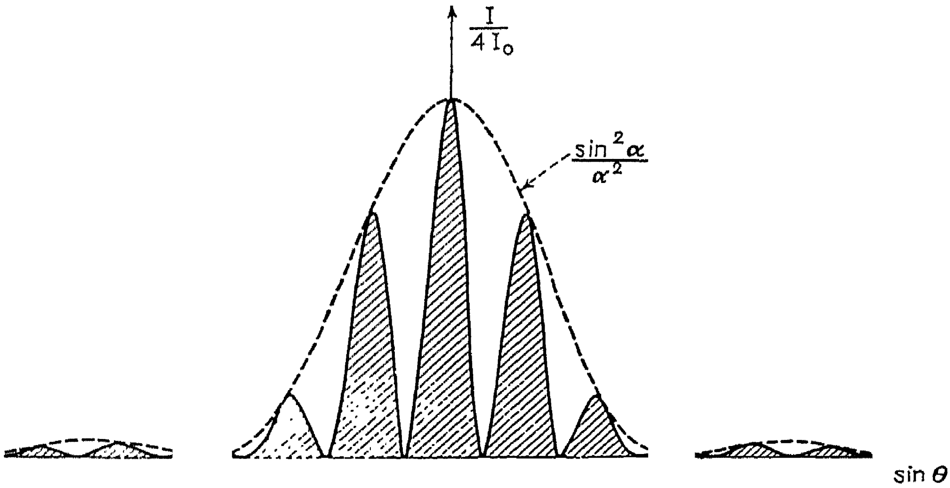

FIG. 228.

the interference pattern of two slits, or, in other words, the diffraction pattern is modulated by the interference pattern. The *minima* occur when either

$$\gamma = (k + \tfrac{1}{2})\pi \qquad (k = 0, 1, 2, 3, \cdots)$$

or when

$$\alpha = (j + 1)\pi \qquad (j = 0, 1, 2, 3, \cdots) \tag{45}$$

Since

$$\alpha = \frac{\pi la}{\lambda} = \frac{\pi a}{\lambda}\sin \qquad \text{and} \qquad \frac{\pi ld}{\lambda} = \frac{\pi d}{\lambda}\sin$$

where θ is the angle which the diffracted rays make with the direction of the incident light, the conditions for minima, expressed by Eq. (45), may be written in the form

$$\left.\begin{aligned} d\sin\theta &= (k + \tfrac{1}{2})\lambda \\ a\sin\theta &= (j + 1)\lambda \end{aligned}\right\} \tag{46}$$

The positions of the *maxima* are not given by any simple relation, but, near the center of the pattern, we may take $\frac{\sin^2 \alpha}{\alpha^2}$ constant. Then the maxima occur approximately at the positions

$$d \sin \theta = k\lambda \tag{47}$$

Figure 228 shows a plot of the intensity curve for a double slit with $d = 3a$.

87. The Diffraction Grating; Many Slits.—In the last section we have seen that the interference effect between two slits produces comparatively sharp maxima in the otherwise slowly varying diffraction pattern of a single slit. This effect can be enhanced tremendously by utilizing a large number N of slits, and in this form one has a plane diffraction grating which is remarkably effective in enabling a spectral analysis of light consisting of a mixture of wave lengths.

Suppose we have a number N of similar parallel slits of width a along the x-axis, with a spacing between centers equal to d. This is the generalization of the double slit problem. The analysis proceeds along exactly the same lines as that of the double slit, and we shall omit the details of the calculation, since it becomes unnecessarily involved when complex numbers are not employed. The result takes a form just like the result for a double slit, *i.e.*, the relative intensity distribution is that of the diffraction pattern of a single slit multiplied by the interference intensity distribution due to a system of N slits. This can be expressed in the formula

$$\frac{I}{I_0} = \left(\frac{\sin^2 \alpha}{\alpha^2}\right) \frac{\sin^2 (\pi N l d/\lambda)}{\sin^2 (\pi l d/\lambda)} \tag{48}$$

Note that this reduces to Eq. (44) for the special case $N = 2$. The factor $\left(\frac{\sin^2 \alpha}{\alpha^2}\right)$ has already been discussed; hence we need only concern ourselves with the second factor in Eq. (48). Different points of the intensity pattern correspond to different values of l, and, as l is varied, both the numerator and denominator of Eq. (48) take on values between zero and unity. The oscillations of the numerator

$$\sin^2 \left(\frac{\pi N l d}{\lambda}\right) \tag{n}$$

give rise to a pattern of interference fringes which are very closely spaced, since the function becomes zero whenever l has values given by

$$l = \sin\theta = j\frac{\lambda}{Nd} \qquad (j = 0, 1, 2, \cdots) \tag{49}$$

In between these minima there will be maxima of the interference fringes which will have different values because of the variations of the denominator, which is

$$\sin^2\left(\frac{\pi l d}{\lambda}\right) \tag{d}$$

and because of the factor $\left(\frac{\sin^2\alpha}{\alpha^2}\right)$. If we disregard the variations of $\sin^2\alpha$ the *least* intense maxima occur when the denominator (d) equals unity, *i.e.*, when

$$\frac{\pi l d}{\lambda} = (j + \tfrac{1}{2})\pi \qquad (j = 0, 1, 2, \cdots) \tag{50}$$

For values of l satisfying Eq. (50), the numerator (n) is also unity, and the intensity of these weakest maxima is given by

$$I = I_0\frac{\sin^2\alpha}{\alpha^2} \cong I_0 \qquad \text{(least intense maxima)} \tag{51}$$

This equation holds very nearly in the region of the center of the diffraction pattern where we may set $\left(\frac{\sin^2\alpha}{\alpha^2}\right)$ equal to unity. This region of validity of Eq. (51) becomes larger, the smaller the individual slit width a.

In the same approximation the most intense maxima occur when the denominator (d) is equal to zero, and these are called the *principal maxima* of the interference pattern. Their location is determined by

$$\sin^2\left(\frac{\pi l d}{\lambda}\right) =$$

or for angles θ given by

$$l = \sin\theta = k\frac{\lambda}{d} \qquad (k = 0, 1, 2, \cdots) \tag{52}$$

Now since we have the relation

$$\lim \text{ of } \frac{\sin^2 (\pi N l d/\lambda)}{\sin^2 (\pi l d/\lambda)} = N^2 \qquad \text{as} \qquad \left(\frac{\pi l d}{\lambda}\right) \to 0$$

we have for the intensities of the principal maxima

$$I = N^2 I_0 \frac{\sin^2 \alpha}{\alpha^2} \qquad \text{(principal maxima)} \tag{53}$$

Thus the ratio between the intensity of the most and least intense maxima is N^2, and this can be made extremely large by making N very large

In a diffraction grating, N is very large, of the order of magnitude of 10^4 or 10^5, so that the secondary maxima are extremely weak compared to the principal maxima. The latter, for monochromatic light, appear then as a series of very sharp lines with intensities given *approximately* by Eq. (53). These lines occur at angles θ with the normal to the grating given by Eq. (52), and this is the ordinary diffraction grating formula, with k determining the so-called *order* of the spectrum. The most intense *secondary* maxima will be adjacent to the principal maxima, and we can compute the relative intensities. The intensity of the kth principal maxima (every Nth maximum is a principal maximum) as given by Eq. (53) occurs for

$$Nl = \frac{kN\lambda}{d}$$

or

$$\frac{lNd}{\lambda} = kN$$

and the neighboring secondary maximum occurs for

$$\frac{lNd}{\lambda} = Nk + \frac{3}{2}$$

Its intensity is very nearly given by

$$I = I_0 \frac{\sin^2 \alpha}{\alpha^2} \cdot \frac{1}{\sin^2 \left(\frac{\pi l d}{\lambda}\right)} = I_0 \frac{\sin^2 \alpha}{\alpha^2} \frac{1}{\sin^2 \left(\pi k + \frac{3\pi}{2N}\right)} \tag{54}$$

and, since N is very large,

$$\sin^2 \left(\pi k + \frac{3\pi}{2N}\right) = \sin^2 \left(\frac{3\pi}{2N}\right) \cong \left(\frac{3\pi}{2N}\right)^2$$

so that Eq. (54) becomes

$$I = N^2 I_0 \frac{\sin^2 \alpha}{\alpha^2}\left(\frac{4}{9\pi^2}\right)$$

with the result that the intensity is $(4/9\pi^2)$ times that of the neighboring principal maximum. This is about 4.5 per cent, and, though small, it is not entirely negligible.

Finally, we must consider the resolving power of such a grating, *i.e.*, its ability to produce separated spectral lines for two almost equal wave lengths. This is determined by the width of the principal maxima, and the customary criterion of resolving power is to consider two spectral lines just distinguishable, if the center of the principal maximum for one lies at the minimum adjacent to the principal maximum of the other. This is called the Rayleigh criterion and is an arbitrary criterion, since the relative intensities of the lines in the incident light play a role as well as the positions of the principal maxima in practice. The position of the center of a principal maximum occurs for $l = k\lambda/d$ according to Eq. (52). This may be written in the form

$$\frac{lNd}{\lambda} = Nk$$

Now the neighboring minimum occurs when

$$\frac{\pi l' Nd}{\lambda} = (Nk + 1)\pi$$

[this makes $\sin^2 (\pi l' Nd/\lambda)$ vanish without having $\sin^2 (\pi l' d/\lambda)$ vanish] or, rewritten,

$$\frac{l' Nd}{\lambda} = Nk + 1$$

Hence the angular separation of these two positions is given by

$$l' - l = \Delta l = \frac{\lambda}{Nd} \tag{55}$$

The condition that two corresponding principal maxima for two different wave lengths λ and $(\lambda + \Delta\lambda)$ have an angular separation Δl is given by Eq. (52) as

$$\Delta l = \frac{k}{d}\Delta\lambda$$

and, substituting in Eq. (55), we find

$$\frac{\lambda}{\Delta\lambda} = Nk \tag{56}$$

This ratio is called the *resolving power* (more precisely, the chromatic resolving power) of the grating and increases both with the order k of the spectrum and with the number of lines N in the grating.

88. Resolving Power of Optical Instruments.—In our study of optical instruments we have entirely neglected diffraction effects, and we must now examine the limitations of these instruments due to these effects, *i.e.*, to the wave nature of light. A lens, for example, will not produce a point image of a point object, even if all aberrations are corrected, since the lens, being of finite cross section, transmits only a limited portion of a wave front incident on it and thus produces a diffraction pattern. An optical system is said to be able to resolve two point objects if the corresponding diffraction patterns are small enough or separated enough to be distinguished as two separate patterns in the image.

Let us consider the case of a telescope objective focused on infinity, and for the moment let us suppose that the lens is square, rather than circular, and of side a. The central diffraction pattern of a point object on the axis of the telescope, such as a very distant star, will subtend a half angle α at the objective given by [compare Eq. (40)]

$$\alpha = \frac{\lambda}{a} \tag{57}$$

Since most of the light falls in this central pattern, we may disregard the presence of the outer diffraction pattern. For a circular lens the computation proceeds along lines similar to that given for the rectangle, but the integration is more difficult. The result turns out to be that the diffraction pattern consists of a central circular disk, on which falls about 85 per cent of the light, surrounded by a series of light and dark rings of rapidly diminishing intensity. The half angle subtended by this central disk is given by

$$\alpha = 1.22\frac{\lambda}{D} = \frac{0.61\lambda}{r} \tag{58}$$

where D is the diameter of the lens and r its radius. Suppose we agree that two point objects can just be resolved if the center of the diffraction disk of one just lies at the periphery of the diffraction disk of the second. One then sees immediately that two stars, for example, will be resolved by a telescope objective if their angular separation β at the objective is equal or greater than α as given by Eq. (58). Thus we must have

$$\beta \geq \frac{1.22\lambda}{D} \tag{59}$$

in order to have resolution of the two images. We see that the larger the diameter of the objective, the greater will be the resolving power.

The eye itself may be considered as a telescope. If we take the diameter of the pupil to be 2 mm., then, for light of wave length 5,500 A. Eq. (59) shows that the minimum angular separation of two point objects just resolvable by the eye is about 1 *minute of arc*.

Let us compute the radius ρ of the central diffraction disk formed by a lens of a point object at a distance l from the lens (Fig. 229). From the figure we see that $\alpha = \rho/l$ and that $\tan\theta_2 = D/2l$, so that, using Eq. (58), we have

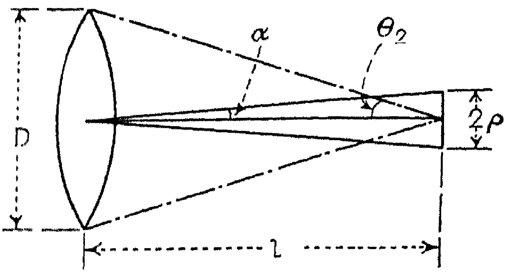

FIG. 229.

$$\frac{\rho}{l} = \frac{0.61\lambda}{l \tan\theta_2}$$

If the medium in which the image is formed has a refractive index n_2, then the wave length λ is related to the wave length λ_0 in air by $\lambda = \lambda_0/n_2$. Thus we may write

$$\rho = \frac{0.61\lambda_0}{n_2 \tan\theta_2} \tag{60}$$

The resolving power of a microscope is conveniently expressed in terms of the linear separation, rather than angular separation, of two point objects which can just be resolved. The distance given by Eq. (60) gives the separation of the centers of the diffraction disks which are the images of two point objects

separated by a distance ρ_0. The object separation ρ_0 is related to the image separation ρ by the magnification of the lens, which, according to Abbe's sine condition [Eq. (14), Chap. XIV], is

$$m = \frac{\rho}{\rho_0} = \frac{n_1 \sin \theta_1}{n_2 \sin \theta_2} \tag{61}$$

where n_1 is the refractive index of the medium in which the object is located and θ_1 is the half angle subtended by the lens *at the object.* Since θ_2 is a small angle (for a microscope objective it is the order of 10^{-2} radian), we may replace $\tan \theta_2$ in Eq. (60) by $\sin \theta_2$ and obtain from Eq. (61)

$$\rho_0 = \frac{0.61\lambda_0}{n_1 \sin \theta_1} \tag{62}$$

The quantity $n_1 \sin \theta_1$ is called the *numerical aperture* of the lens. The larger the numerical aperture, the better the resolving power of the lens. For air the upper limit is about 0.95 in practice, but, by immersing the object in oil as is done in high-powered microscopes, the numerical aperture may be increased to 1.60. Physically, the effect of the oil is to enable one to use waves of shorter wave length than in air.

The numerical aperture of the unaided eye, using a pupillary radius of 1 mm. and an object distance in air of 25 cm., is

$$\text{N.A.} = n_1 \sin \theta_1 = 1 \cdot \tfrac{1}{250} = 0.004$$

so that the smallest separation of two points just distinguishable as separate objects at a distance of 25 cm. is, according to Eq. (62),

$$\rho_0 = \frac{0.61 \times 5 \times 10^{-5}}{4 \times 10^{-3}} = 0.075 \text{ mm.}$$

for light of wave length 5,000 A.

On the other hand, for a microscope objective with a numerical aperture of 1.60, this distance is smaller in the ratio

$$\frac{1.6}{0.004} = 400,$$

so that it is about 2×10^{-5} cm., about half a wave length of light. Since the microscope gives 400 times the resolving power of the naked eye, the magnifying power should be at least 400 times to take advantage of this. *The ratio of the numerical*

aperture of a microscope objective to that of the eye is called the normal magnifying power of a microscope. A lower magnifying power will not take full advantage of the resolving power available, and a larger magnifying power gains nothing in detail and loses in brightness of the image. However, if enough light is available, one frequently employs higher magnifying powers for ease of observation. One defines the normal magnifying power of a telescope in a similar manner.

Problems

1. Two plane waves, each of wave length 6,000 A., travel at an angle θ as shown in Fig. 213 and form interference fringes. If the smallest separation between neighboring light and dark fringes is to be 0.01 mm., what is the largest angle θ allowable if the fringes are to be observed?

2. In Young's experiment the separation of the slits is 0.1 mm., and the distance from slits to screen is 1 meter. Compute the distance between neighboring dark fringes:

a. For blue light of wave length 4,000 A.

b. For red light of wave length 7,000 A.

3. Plot the relative intensity of light on the screen in Young's experiment as a function of x, the coordinate of the point P of Fig. 215. Prove that the average intensity is the same as would exist in the absence of interference, so that no energy is lost in the process of interference.

4. Suppose that the light employed in a Young's experiment consists of a mixture of two wave lengths λ_1 and λ_2, almost equal to each other. Derive an expression for the difference of these wave lengths such that one of the maxima for one wave length is located at the position of a neighboring intensity minimum of the other.

5. Find the thickness of a plane soap film of refractive index 1.33 for a strong first-order reflection of the red hydrogen line of wave length 6,563 A. at normal incidence. What is the wave length of the light inside the film?

6. Two pieces of plane plate glass are placed together with a piece of paper between the two at one edge. When viewed at normal incidence with sodium light ($\lambda = 5{,}893$ A.), eight interference fringes per centimeter are observed. Find the angle of the wedge-shaped air film between the plates.

7. If the radius of curvature of the convex surface of the plano-convex lens used in producing Newton's rings is 5 meters, what will be the diameters of the fifth and tenth bright rings for the red hydrogen line, $\lambda = 6{,}563$ A.?

8. Newton's ring experiment is performed with violet light using a convex lens surface of radius of 10 meters. The radius of the kth dark fringe is 4.0 mm. and that of the $(k + 5)$th dark fringe is 6.0 mm. Find the wave length of the light used and the ring number k.

9. Prove that the sum of the amplitudes of all the reflected rays excepting the first in Fig. 216 is equal to the amplitude of the first reflected ray. Assume normal incidence.

10. A so-called "zone plate" is constructed by constructing 20 Fresnel zones of the type discussed in Fig. 222 and by blocking off the light from every *other* zone. This zone plate, when held in the light from a distant point source, produces a bright spot on its axis 100 cm. from the plate. Assuming a wave length of 5,000 A. compute:

a. The areas of the zones on the plate.

b. The radius of the zone plate.

c. The intensity of light at the bright spot relative to its value at this point in the absence of the zone plate.

11. Plot a graph of the intensity of light at the point P of Fig. 224 as a function of distance D from a fixed circular aperture. A qualitative result is all that is wanted.

How would this graph be altered if the circular aperture were replaced by an opaque circular obstacle of the same radius?

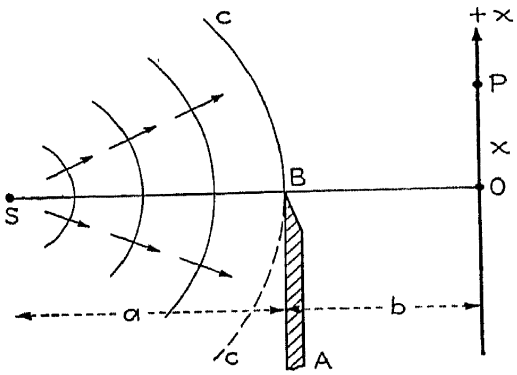

FIG. 230.

12. Consider light from a *line* source S perpendicular to the plane of the paper in Fig. 230, passing a straightedged obstacle B to a screen, as shown, the wave fronts being cylindrical. Show that the Fresnel zones are strips on the wave front *cc*. Find expressions for the angles subtended by these zones at the source S.

13. In the experiment of Fig. 230 discuss qualitatively the variation of intensity on the screen as a function of x, the coordinate of P, when the obstacle AB is in the position shown. Show that the diffraction pattern consists of a series of parallel light and dark fringes, the maxima being located at values of x (positive) given approximately by

$$= \sqrt{\frac{b}{a}(a + b)\lambda(2k + 1)}$$

and the minima at

$$= \sqrt{\frac{b}{a}(a + b)2k\lambda}$$

Make a rough plot of intensity as a function of x for both positive and negative values of x.

14. One wishes to observe the bright spot at the center of the shadow of a circular disk of radius 2.0 mm. which is placed in the beam of light coming from a point source far enough away so that the incident light on the disk may be taken as a plane wave. For a wave length of 6,500 A., how far must the screen be from the disk so that the first two Fresnel zones are covered by the disk?

15. A single slit Fraunhofer diffraction pattern is formed with white light. For what wave length of light does the third maximum in its diffrac-

tion pattern coincide with the second maximum in the pattern for red light of wave length 6,500 A.?

16. A plane light wave of wave length 5,000 A. is normally incident on a slit 1 mm. wide and 4 mm. long. A lens of focal length 100 cm. is mounted just behind the slit and the light focused on a screen. Find the dimensions of the central part of the diffraction pattern in millimeters.

17. In the double-slit pattern show that, if the slit separation is twice the width of either slit, all the even-order interference maxima will be missing. What orders will be missing if the ratio is 3:1?

18. Compute the relative intensities of the first five principal maxima of a double-slit diffraction pattern for which $d = 5a$. Sketch the intensity distribution for a sufficiently large angular range to include these five maxima.

19. Plot the intensity distribution in the diffraction pattern formed by a grating of four equally spaced slits with $d = 3a$.

20. The limits of the visible spectrum are nearly 4,000 to 7,000 A. Find the angular breadth of the first-order visible spectrum formed by a plane grating with 12,000 lines per inch.

Does the violet of the third-order visible spectrum overlap the red in the second-order spectrum? If so, by how much (approximately)?

21. Light containing two wave lengths of 5,000 and 5,200 A. is normally incident on a plane diffraction grating having a grating spacing of 10^{-3} cm. If a 2-meter lens is used to focus the spectrum on a screen, find the distance between these two lines (in centimeters) on the screen:

a. For the first-order spectrum.

b. For the third-order spectrum.

22. The sodium yellow line 5,893 A. is a doublet with a separation of 6 A. between the two lines. What is the minimum number of lines of a grating which will just resolve these lines in the third-order spectrum?

23. Compute the approximate radius of the central diffraction disk formed on the retina of the eye by a distant point object, assuming a pupillary diameter of 2.0 mm. The distance from the cornea to the retina is about 1 in., and the index of refraction of the vitreous humor, the medium in which the image is formed, is 1.33.

24. Find the angular separation in seconds of arc of the closest double star which can be resolved by the 40-in. diameter Yerkes refracting telescope.

25. Two pinholes, 1 mm. apart, are made in a screen and placed in front of a bright source of light. They are viewed through a telescope with its objective stopped down to a diameter of 1 cm. How far from the telescope may the screen be and still have the pinholes appear as separate sources for a wave length of 5,000 A.?

26. An oil-immersion microscope will just resolve a set of test lines drawn 112,000 to the inch, using blue light of wave length 4,200 A. Find the numerical aperture of the objective.

27. Ultraviolet light of wave length 2,750 A. is used in photomicrography in conjunction with a quartz-lens microscope. Assuming a numerical aperture of 0.85, what is the smallest separation of two points which can be resolved?

28. If the focal length of a microscope objective is 5.00 mm. and its numerical aperture is 0.85, what focal-length ocular should be used? What is the smallest separation of two objects just resolvable with this microscope?

29. Prove that the maxima of the function $(\sin^2 \alpha)/\alpha^2$ occur at values of α given by

$$\alpha = \tan \alpha$$

Find the first three roots of this equation (excluding $\alpha = 0$), compute the corresponding values of the function, and compare with the approximate values $(2/3\pi)^2$; $(2/5\pi)^2$ and $(2/7\pi)^2$, obtained by taking the maxima at positions halfway between the minima.

CHAPTER XVII

HEAT RADIATION

We have repeatedly stressed the fact that the radiation emitted by material bodies as heat, light, X rays, etc., is due to the combined effects of many molecules and atoms and is incoherent. The nature of this radiation depends, in general, on the mode of excitation of the emitting atoms and molecules and on their specific properties. By the nature of the radiation we mean its spectral distribution, polarization, intensity, etc. In this chapter we shall concern ourselves principally with that form of radiation known as *heat* or *thermal radiation* because of the mode of excitation. It is a familiar fact that material bodies begin to emit invisible heat radiation (infrared waves) as they are heated; then, as the temperature is increased, they emit visible radiation with increasing intensity in the short wave-length region. Furthermore, the total rate of emission of energy increases very rapidly with increasing temperature.

The transfer of energy by thermal radiation is a process which differs fundamentally from the corresponding transfer of energy by thermal conduction. In the latter case one can describe uniquely the process by a single vector (the heat current density) at each point of the medium, and this vector depends on the local temperature gradient. On the other hand, thermal radiation at a given point of space or of a material body cannot be represented by a single vector and does not in general depend on the temperature or temperature gradient at the point in question. In fact, it is necessary to employ the concept of an infinite number of rays passing through a point in all conceivable directions to describe the radiation state, and these rays are all mutually independent with regard to their intensities, frequencies, and polarizations. Even two rays of equal frequency and polarization and opposite directions of propagation do not combine to form a single ray but maintain their individual identities.

89. Emission and Absorption; Kirchhoff's Law.—In building appropriate definitions to describe the state of radiation, we must keep in mind the fact that radiation of finite energy content can never be emitted from point sources but must come from bodies of finite size, and, inasmuch as the radiation emerges through the surface of a radiating body, one may say that radiation always comes from, impinges on, or passes through an element of surface but not a point. We shall, in our study, employ the approximation of geometrical optics, so that we may follow the propagation of energy in terms of bundles of rays. This implies that we may choose elementary areas large compared to the wave lengths under consideration ($dA >> \lambda^2$) but very small compared to ordinary dimensions. Furthermore, it should be pointed out that one can never realize a beam of strictly parallel rays, but the bundle of rays must form a converging or diverging cone of given direction and small solid angle.

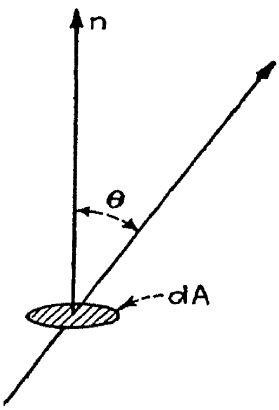

FIG. 231.

Let us consider a material medium or a region of empty space which is being traversed by radiation and focus our attention on an elementary area dA at a given point. This area will be traversed by rays propagating in all directions, and the energy crossing this area per unit time in a direction making an angle θ with the normal to this surface (Fig. 231) will evidently be proportional to $dA \cos \theta$, *i.e.*, to the projection of dA on a plane which is perpendicular to the chosen direction. The bundle of rays traversing dA in this direction forms a small cone or narrow pencil with vertex at dA, and the energy flow from or to dA per unit time will be proportional to the solid angle $d\omega$ at the vertex of this cone (Fig. 232). Thus we may write for the energy per unit time (the energy flux dF) crossing dA in a bundle of rays subtending a solid angle $d\omega$ at dA, when the bundle has a direction θ with the normal,

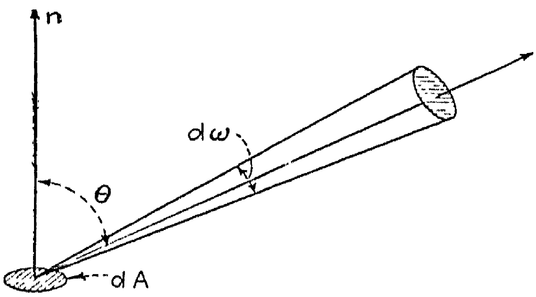

FIG. 232.

$$dF = K \cos \theta \, dA \, d\omega \tag{1}$$

The proportionality constant K is called the *specific intensity* or *brightness* of this pencil of rays. This brightness K may be further resolved into the specific intensities of those rays in the bundle corresponding to different frequencies or wave lengths in the radiation and hence has a definite spectral composition. If we consider those rays, the frequencies of which lie between ν and $\nu + d\nu$, we may write as the contribution to K from this spectral range $d\nu$,

$$K = K_\nu \, d\nu$$

so that

$$K = \int_0^\infty K_\nu \, d\nu \tag{2}$$

Now let us consider the nature of the radiation which is in thermal equilibrium with a material body. For this purpose consider an evacuated enclosure or cavity of arbitrary shape, and let us suppose that we have attained thermal equilibrium. The walls of the enclosure will all be at the same temperature T, and they will be constantly emitting and absorbing radiation. If the walls are constructed of material which absorbs *all* the radiation incident on them and reflects none, we say that they are *black*. The radiation in the cavity will then be isotropic and homogeneous, the specific intensity K will be independent of position and direction at any point in the radiation field, and there will be no preferred state of polarization. The *energy density* of the radiation, which has the same value at every point, will depend only on the temperature of the walls, and there will be a definite spectral distribution of this energy density at each temperature. Under these conditions we say that the radiation is *black*, or that it is *black-body radiation.*

There is a simple relation between the brightness K and the energy density u of radiation at a given point for the case of isotropic radiation. To obtain this relation, we must consider the energy arriving at the point under consideration coming from all directions. Consider the energy arriving at an elementary area dA in a direction θ with the normal to dA. In time dt the energy transported across dA by these rays fills an infinitesimal cylinder of slant height $c\,dt$ (Fig. 233), where c is the velocity of propagation, and of base dA, so that it fills a volume equal to $c \cos \theta \, dt \, dA$. According to Eq. (1), this energy is equal to

$$dF \cdot dt = K \cos\theta \, dA \, d\omega \, dt$$

Hence this bundle yields a contribution du to the energy density equal to

$$du = \frac{K \cos\theta \, dA \, d\omega \, dt}{c \cos\theta \, dt \, dA} = \frac{1}{c} K \, d\omega \tag{3}$$

The total energy density u is the sum of the contributions from all the pencils crossing dA in *all* directions, and this is evidently obtained by integrating Eq. (3) over all these directions. Hence we may write

$$u = \frac{1}{c} \int K \, d\omega$$

and, since for isotropic radiation K is independent of direction, this becomes

$$= \frac{K}{c} \int d\omega = \frac{4\pi}{c} K \tag{4}$$

using the fact that the solid angle encompassing all directions at a point is 4π.

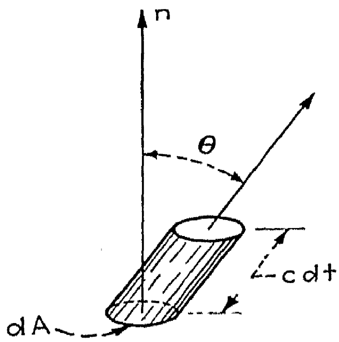

FIG. 233.

Now consider the case for which the walls of our enclosure are perfectly reflecting surfaces (diffuse reflectors) for all wave lengths, so that no energy is absorbed. In this case the radiation in the cavity may have any composition whatsoever, since the various rays do not interact with each other and there is no mechanism by which the existing spectral distribution or the state of polarization may be altered. If we introduce a tiny black body which is at the same temperature as the walls, it will absorb and reemit radiation so that, after thermal equilibrium for the whole system is established, we again have black-body radiation at temperature T. Since the black body so introduced may be made as small as we please, its contribution to the energy of the system may be disregarded, and it acts simply as a catalyst which insures the black-body distribution and composition of the radiation in the cavity. The black-body radiation in a perfectly reflecting enclosure (with a speck of black body at temperature T) may be said to have a temperature T equal to that of the black body with which it is in equilibrium,

since its space distribution and spectral composition are determined uniquely by this temperature. If we place an arbitrary body (not black) of the *same temperature* in the cavity, the state of the radiation must remain unchanged (thermal equilibrium). Now the total rate of energy flow across a closed surface surrounding this body must be zero; hence

$$F = \int_{\substack{\text{closed}\\ \text{surface}}} S_n \, dA = 0$$

where S_n is the normal component of the Poynting vector at dA. F is composed of the incident radiation F_0 which enters the surface from outside; the reflected radiation F_r; the radiation F_e emitted by the body; and finally the radiation F_t which is transmitted through the body and emerges through the other side of the surface. Hence we must have

$$F_0 = F_r + F_e + F_t \tag{5}$$

In traversing the body, a fraction a of the incident radiation is absorbed. a is called the *absorption power* of the body. Evidently for the radiation absorbed, we must have

$$aF_0 = F_0 - F_r - F_t \tag{6}$$

From Eqs. (5) and (6) there follows immediately

$$aF_0 = F_e$$

or, dividing by A, the area of the body,

$$E = \frac{F_e}{A} = a\frac{F_0}{A} \tag{7}$$

E is the *emissive power* of the body at temperature T, and we see that it is greatest when $a = 1$, *i.e.*, for a black body which absorbs all incident radiation. For this reason, a black body is often referred to as an ideal radiator, and it yields the maximum thermal radiation which can be obtained at a given temperature. For such a black body, Eq. (7) becomes

$$\bar{E} = \frac{F_0}{A} \tag{8}$$

where $\bar{E}$ is the emissive power of the black body. Using this relation, Eq. (7) may be written as

$$E = a\bar{E} \tag{9}$$

The above relations hold, not only for the total radiation of all frequencies, but also for that portion of the radiation in the spectral range $d\nu$, *i.e.*, for frequencies lying between ν and $\nu + d\nu$. We have thus obtained a fundamental result known as *Kirchhoff's law: The emissive power of a body is equal to its absorption power multiplied by the emissive power of a black body at the same temperature.* It simplifies the discussion of radiation considerably, since the emission of radiation by any material can be referred to that emitted by an ideal black body, and the properties of the body which one requires are simply its absorption power. If a body is transparent for any range of frequencies or wave lengths, it cannot radiate energy of these wave lengths. The ratio of the emissive power E of a body to that of a black body $\bar{E}$ is termed the emissivity ϵ. It is equal to the absorbing power a of the body.

Finally, let us consider the dependence of the emissive power of a surface element on the angle θ which the emitted bundle of rays makes with the normal to the surface. The energy incident on the area element dA per unit time in this bundle of angular opening $d\omega$ is given by Eq. (1), so that we have

$$dF_0 = K \cos\theta \, dA \, d\omega$$

If dA is an element of area of a black body, all this incident energy will be absorbed, and, if the radiation state is not disturbed (thermal equilibrium), the body must emit a similar bundle; hence

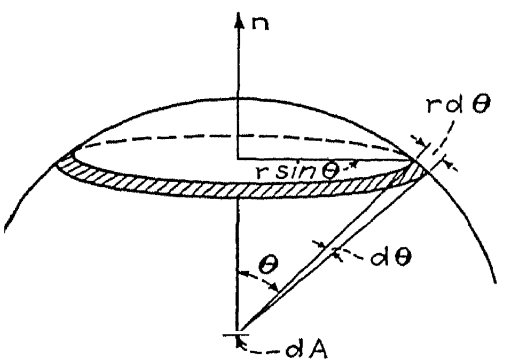

FIG. 234.

$$dF_e = K \cos\theta \, dA \, d\omega \tag{10}$$

The specific intensity or brightness K is that corresponding to black-body radiation and is related to the energy density thereof by Eq. (4). To calculate the total emission rate from this surface element, consider the radiation emitted in the hollow conical bundle between θ and $\theta + d\theta$ (Fig. 234). The solid angle of this hollow cone is $d\omega = 2\pi \sin\theta \, d\theta$, since, on a sphere of radius r, the area of the ring is

$$2\pi r^2 \sin\theta\, d\theta$$

and by definition $d\omega$ is this area divided by r^2. Thus from Eq. (10) we have

$$F_e = 2\pi K\, dA \int_0^{\frac{\pi}{2}} \cos\theta \sin\theta\, d\theta \tag{11}$$

as the emission rate from one side of the surface element dA (the outside). The value of the integral is $\frac{1}{2}$, so that we obtain

$$F_e = \pi K\, dA \tag{12}$$

The energy radiated per unit time per unit area is the emissive power $\bar{E}$ of the surface, so that

$$\bar{E} = \pi K \tag{13}$$

Equation (13), when applied to sources emitting visible radiation, is known as *Lambert's law*, and, although we have shown it to be true for black bodies, it turns out experimentally to be very nearly true for some sources which are not black. When applied to such cases, the brightness K of the source will in general be different from that of a black body.

90. Radiation Pressure.—When an electromagnetic wave impinges on the surface of a material body, it exerts a mechanical force on the body in the direction of propagation of the wave. In general, one must deal with both normal stresses (pressures) and shearing stresses on the surfaces of bodies on which radiation is incident (or from which radiation is emitted). For the case of normal incidence, or for the case of isotropic radiation which is fundamental in our study of heat radiation, one has to do with the pressure of radiation, and, before proceeding farther with the question of the laws of emission of thermal radiation, we must derive the relation between radiation pressure and the energy density of the radiation exerting the pressure.

Let us start with the simple case of a plane electromagnetic wave normally incident on the surface of an ideal metal of infinite conductivity. Inside the metal there can be no electric field and hence no magnetic field nor electromagnetic wave; therefore the metal is a perfect reflector. There will be a surface current of surface density j induced on the conductor surface, as shown in Fig. 235, and this alternating current emits the reflected wave. The magnitude of j will be such that its magnetic field

inside the metal just cancels that of the incoming wave and doubles the magnetic field just outside the conductor surface. The electric vectors of the incident and reflected waves must be equal and opposite at each instant of time just outside the metal surface, since the tangential components of $\mathcal{E}$ are continuous at any boundary. The mechanical force on the induced surface current exerted by the magnetic field of the incident wave is directed into the metal, as is evident from Fig. 235, and this is the origin of the pressure of radiation for this case.

Consider an element of area of width dw and length ds as shown in Fig. 235. The current flowing on this elementary area forms a current element $i\,ds = j\,dw\,ds = j\,dA$, and, since the magnitude of the force on this current element is given by

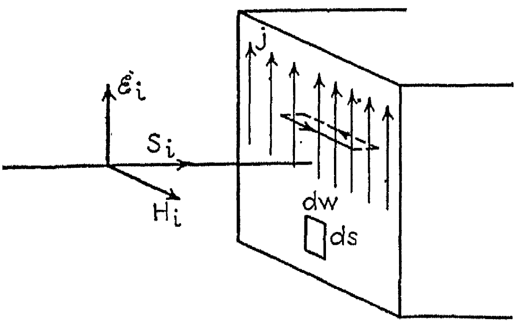

FIG. 235.

$$dF = i\,ds\,B_i = jB_i\,dA$$

we have, for the pressure p,

$$p = jB_i = \mu_0 jH_i \qquad (14)$$

where B_i is the magnetic induction vector of the incident wave just at the metal surface. There remains the task of computing the surface-current density j, and this can be readily accomplished with the help of the Ampère circuital law. The lines of H produced by the surface current run parallel to the surface at right angles to j, and they are in opposite directions just outside and inside the surface (compare Prob. 40, Chap. V). Hence we employ the closed path shown in Fig. 235 to compute the m.m.f. and find for the magnitude of the magnetic intensity produced by the surface current on either side of the surface

$$H_r = 2\pi j \qquad (15)$$

where we have written H_r to indicate that this is the magnetic vector of the reflected wave just at the surface. Inside the metal we have H_r equal and opposite to H_i, so that no wave exists in the metal, whereas just outside the surface we have for the resultant magnetic intensity

$$H = H_i + H_r = 2H_i = 2H_r \qquad (16)$$

Using Eqs. (15) and (16) in Eq. (14), we then find

$$p = \frac{\mu_0 H^2}{8\pi} \tag{17}$$

and, since the electric energy density is zero just at the conductor surface, we may write

$$p = u \tag{18}$$

where u is the electromagnetic energy density just at the surface of the reflector. It is interesting and instructive to consider the meaning of this result in terms of the energy densities of the incident and reflected waves. The incident wave has an energy density given by

$$u_i = \frac{1}{8\pi}\left(\epsilon_0 \mathcal{E}_i^2 + \mu_0 H_i^2\right) = \frac{\mu_0 H_i^2}{4\pi} = \frac{S_i}{c} \tag{19}$$

where S_i is the magnitude of the Poynting vector, and we have used the fact that for a plane wave $\epsilon_0 \mathcal{E}^2 = \mu_0 H^2$. Similarly, for the reflected wave

$$u_r = \frac{1}{8\pi}\left(\epsilon_0 \mathcal{E}_r^2 + \mu_0 H_r^2\right) = \frac{\mu_0 H_r^2}{4\pi} = \frac{S_r}{c} \tag{20}$$

According to Eq. (16), $H_i = H_r$ at the conductor surface; hence we may write for the pressure as given by Eq. (17),

$$p = \frac{\mu_0 H_i^2}{2\pi} = \frac{\mu_0 H_r^2}{2\pi}$$

or

$$p = \frac{\mu_0 H_i^2}{4\pi} + \frac{\mu_0 H_r^2}{4\pi} = u_i + u_r \tag{21}$$

so that the radiation pressure on the metal equals the sum of the energy densities of the incident and reflected waves at its surface.

Thus we are led to the conclusion that electromagnetic waves transport not only energy (the flow given by the Poynting vector) but also momentum. The momentum carried by these waves may be thought of as distributed throughout space in a manner similar to the energy; hence we may introduce the idea of a space density of electromagnetic momentum g. Consider an element of area ΔA of the conductor surface. In time dt, the momentum incident on this surface is given by

$$g_i \, \Delta A c \, dt$$

and the momentum carried away from this surface element by the reflected wave in time dt is similarly given by

$$g_r \, \Delta A c \, dt$$

where $g_i = g_r$ for the case under consideration. The change of momentum in time dt is the sum of these two expressions, and consequently the force on the area ΔA is, by Newton's second law,

$$\Delta F = c(g_i + g_r) \, \Delta A$$

and the pressure

$$p = cg_i + cg_r \tag{22}$$

Comparing this with Eq. (21), we see that the momentum density in an electromagnetic wave is related to the energy density, and hence to the Poynting vector, by

$$g = \frac{u}{c} = \frac{S}{c^2} \tag{23}$$

More generally, the vector relation between $\overrightarrow{g}$ and $\overrightarrow{S}$ is

$$\overrightarrow{g} = \frac{1}{c^2}\overrightarrow{S} \tag{23a}$$

The concept of electromagnetic momentum now enables us to compute the pressure of isotropic radiation. Before doing so, however, we must emphasize the fact that the relation given by Eq. (18) for normal incidence is true, not only for perfect reflectors, but for arbitrary surfaces, if by u we mean the *total* electromagnetic energy density at the surface. For example, let us suppose that we have a perfectly absorbing surface, so that there is no reflection, and that this surface is at such a low temperature that it emits negligible radiation. Then Eqs. (22) and (23) show that the pressure of normally incident radiation is still equal to the energy density at the surface ($p = u_i$ and $u_r = 0$), but, since the energy density is now that of the incident wave alone which, according to Eqs. (21) and (16), is half as large as for the case of perfect reflection, the pressure is reduced by half also. Now let us consider radiation impinging on an element dA of the surface of a body at an angle θ with the normal. The momentum transferred to the surface element in time Δt is that contained in a slant prism of

base ΔA and slant height $c\ \Delta t$ (Fig. 236). This gives rise to a force ΔF in the direction shown given by

$$\Delta F = gc \cos\theta\ \Delta A$$

or, since by Eq. (23) $g = u/c$, this can be written as

$$\Delta F = u \cos\theta\ \Delta A \tag{24}$$

This gives rise to a component normal to the surface

$$\Delta F_n = u \cos^2\theta\ \Delta A \tag{25}$$

and a tangential component

$$\Delta F_t = u \cos\theta \sin\theta\ \Delta A \tag{26}$$

For isotropic radiation (u independent of direction) the tangential components of ΔF as given by Eq. (26) sum up to zero, and the

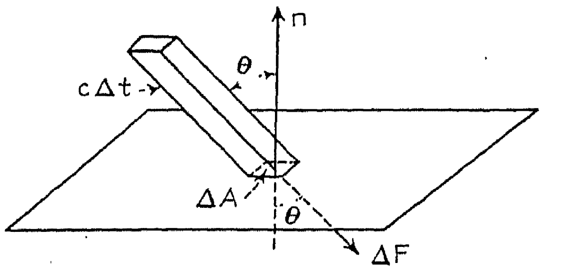

FIG. 236.

pressure is obtained from Eq. (25). Since the average value of $\cos^2\theta$ over a hemisphere is $\frac{1}{3}$, this yields for the *pressure of isotropic radiation*

$$p = u \tag{27}$$

This is a fundamental equation in radiation theory. One more remark may be appropriate at this point. Let us suppose that we have a hollow enclosure containing black-body radiation at a temperature T, and let the temperature of the walls (which may be of arbitrary composition) be T, so that the system is in thermal equilibrium. From the second law of thermodynamics we must also have mechanical equilibrium for this system, since otherwise mechanical work could be obtained at the expense of the internal energy of an isolated system, all parts of which are at the same temperature. It thus follows that the pressure must be the same at all points of the walls, independent of

their absorbing or reflecting powers, and is related to the energy density by Eq. (27).

91. The Stefan-Boltzmann Law.—We have already pointed out that the energy density of black-body radiation depends only on its temperature. The law expressing this dependence was found experimentally by *Stefan* and later deduced theoretically by *Boltzmann*. This law states that the emissive power of a black body is proportional to the fourth power of its absolute temperature. If, as before, $\bar{E}$ denotes the emissive power of a black body, we can write

$$\bar{E} = \sigma T^4 \tag{28}$$

where σ, the Stefan-Boltzmann constant, has the value

$$\sigma = 5.74 \times 10^{-5} \text{ erg/cm.}^2\text{-sec.-°C.}^4 = 1.37 \times 10^{-12} \text{ cal./cm.}^2\text{- sec.-°C.}^4 \tag{28a}$$

We have written Eq. (28) for the emissive power $\bar{E}$, but we can readily see that a similar expression holds for the radiation density u. Using Eq. (13) to express $\bar{E}$ in terms of K, and Eq. (4) to express K in terms of u, we find readily that

$$u = \frac{4\pi}{c}K = \frac{4}{c}\bar{E} = \frac{4\sigma}{c}T^4 = \alpha T^4 \tag{29}$$

The thermal radiation from many real surfaces which are not black is found experimentally to be very nearly proportional to the fourth power of the absolute temperature, but with a proportionality constant which is smaller than σ as given by Eq. (28a). This is the case for metals such as platinum and tungsten and also for carbon. For such surfaces, one can write

$$E = \epsilon\sigma T^4 \tag{30}$$

where ϵ, the emissivity, has already been referred to and is equal to the absorbing power of the body (Eq. 9). The emissivity of a hot tungsten-lamp filament is about $\frac{1}{3}$.

The Stefan-Boltzmann law can be derived theoretically from the second law of thermodynamics in the following manner: Consider a cylinder with a tightly fitting frictionless piston and perfectly diffuse reflecting walls which contains black-body radiation at temperature T. Let the volume in which the radiation is present be V and the walls be at the temperature T

(Fig. 237). As we have shown, the radiation exerts a pressure $p = \frac{1}{3}u$ on the piston and the energy density u is uniformly distributed throughout the volume. The internal energy of the system is uV, where u depends only on the temperature T. Now let an amount of heat dQ flow reversibly into the system. The first law of thermodynamics requires that

$$d(uV) = dQ - p\,dV$$

where dV is the volume change corresponding to a small motion of the piston. Since

$$d(uV) = u\,dV + V\,du$$

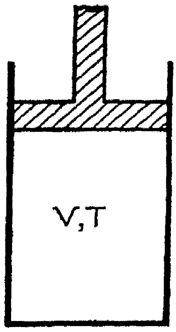

FIG. 237.

the above equation may be written in the form

$$dQ = V\,du + (p + u)\,dV$$

and since $p = u/3$, this takes the form

$$dQ = V\left(\frac{du}{dT}\right) dT + \frac{4}{3}u\,dV \tag{31}$$

where we have used the fact that u is a function of T only and not of V.

Now the second law of thermodynamics requires that the heat dQ added in a reversible process divided by the absolute temperature T (the change of entropy) depend only on the initial and final states of the system and not on the intermediate stages of the process. This is equivalent to saying that, if we divide each term of Eq. (31) by T, the left-hand side becomes a total differential. Thus in the equation

$$dS = \frac{dQ}{T} = \frac{V}{T}\left(\frac{du}{dT}\right) dT + \frac{4}{3}\frac{u}{T}\,dV \tag{32}$$

$\frac{V}{T}\left(\frac{du}{dT}\right)$, the coefficient of dT, is $\left(\frac{\partial S}{\partial T}\right)_V$, and $\frac{4}{3}\frac{u}{T}$, the coefficient of dV, is $\left(\frac{\partial S}{\partial V}\right)_T$. Since, in general, the rules of partial differentiation demand that

$$\frac{\partial^2 S}{\partial T\,\partial V} = \frac{\partial^2 S}{\partial V\,\partial T}$$

there follows

$$\frac{\partial}{\partial V}\left(\frac{V}{T}\frac{du}{dT}\right) = \frac{\partial}{\partial T}\left(\frac{4}{3}\frac{u}{T}\right)$$

Remembering that u depends only on T, this becomes

$$\frac{1}{T}\frac{du}{dT} \qquad \frac{4}{3T}\frac{du}{dT} \qquad \frac{4}{3}\frac{u}{T^2}$$

or

$$\frac{du}{dT} = 4\frac{u}{T}$$

Separating variables,

$$\frac{du}{u} \qquad 4\frac{dT}{T}$$

and integrating

$$\ln u = \ln T^4 + \text{constant}$$

or

$$u = \alpha T^4 \tag{33}$$

which is identical with Eq. (29).

92. The Planck Radiation Law; Wien's Displacement Law.—There still remains the fundamental question of the spectral composition of black-body radiation. The Stefan-Boltzmann law gives the temperature dependence of the total energy density u, but places no limitations whatsoever on the possible spectral distribution of this energy density among the various wave lengths or frequencies of the electromagnetic spectrum. Indeed, thermodynamics alone cannot provide a unique answer to this question. The situation is somewhat analogous to the theory of gases. Here thermodynamics can provide us with a number of relations involving the internal energy of the gas, but the velocity distribution of the molecules is a matter for atomic theory. The researches of Planck, about 1900, in connection with this problem led to the foundations of the now famous quantum theory. Classical physics led to an impossible law, as we shall point out shortly. It would be far beyond the scope of this book to attempt a discussion of the quantum theory, and we must content ourselves with a statement and discussion of the results.

If we denote the energy density per unit frequency range of the spectrum by u_ν, then in the frequency range $d\nu$ there will

be a contribution to the total energy density equal to $u_\nu\, d\nu$, and consequently

$$u = \int_0^\infty u_\nu\, d\nu \tag{34}$$

As we have seen in connection with the Stefan-Boltzmann law, the brightness K and the emissive power of a black body $\bar{E}$ are proportional to u. Similarly a proportionality exists between these quantities per unit frequency or per unit wave-length range. Although it is more convenient from a theoretical standpoint to deal with frequency distribution, it is more convenient experimentally to deal with the distribution with respect to wave length, and we must say a word or two concerning these two modes of description. For the sake of concreteness, consider the emissive power of a black body. This can be written as

$$\bar{E} = \int_0^\infty E_\lambda\, d\lambda = \int_0^\infty E_\nu\, d\nu$$

where E_λ and E_ν are the emissive powers per unit wave length and frequency range, respectively. The spectral distribution of black-body radiation can be specified by giving either E_λ as a function of λ or by giving E_ν as a function of ν. The resulting formulas will *not* be identical in form, as we can see from the following: Consider a range of frequencies between ν and $\nu + d\nu$ and a corresponding range of wave lengths $d\lambda$. The energy of these wave lengths radiated per unit time per unit area by a black body is given by

$$d\bar{E} = E_\nu\, d\nu = E_\lambda\, d\lambda$$

and, since the relation between frequency and wave length is

we have immediately

$$|d\nu| = \frac{c}{\lambda^2}|d\lambda|$$

so that E_ν and E_λ are related by the equation

$$E_\lambda = \frac{c}{\lambda^2}E_\nu \tag{35}$$

Similarly for energy densities, we have $u_\lambda = \dfrac{c}{\lambda^2}u_\nu$. Equation

(35) must be kept in mind when translating experimental results from a wave length to a frequency scale. Figure 238 shows the shape of the curves for E_λ for black-body radiation as a function of wave length, as obtained from experiment. In these curves $T_1 > T_2 > T_3$. Note that the maxima of these curves moves toward shorter wave lengths as the temperature is increased, as one would expect from the color changes which occur when the temperature of a body is raised. The curves of Fig. 238 may be looked upon as plots of u_λ as function of λ, since there would only be a difference of a scale factor.

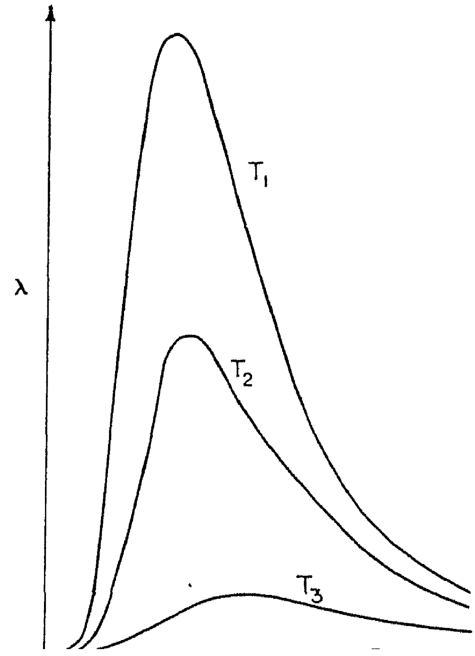

FIG. 238.

The equation describing the normal spectrum (black-body spectrum) obtained by Planck is as follows:

$$u_\nu = \frac{8\pi h\nu^3}{c^3}\,\frac{1}{e^{\frac{h\nu}{kT}} - 1} \tag{36}$$

Here h is a fundamental atomic constant known as Planck's constant and has the value

$$h = 6.55 \times 10^{-27} \text{ erg-sec.}$$

k is Boltzmann's constant (1.37×10^{-16} erg/°C.) and T the absolute temperature. For the wave-length distribution we have, using the fact expressed by Eq. (35),

$$u_\lambda = \frac{8\pi hc}{\lambda^5}\,\frac{1}{e^{\frac{hc}{\lambda kT}} - 1} = \frac{c_1/\lambda^5}{e^{\frac{c_2}{\lambda T}} - 1} \tag{37}$$

which is the form usually employed by experimentalists. There are two limiting cases for which we may rewrite Eq. (36) which have interesting historical significance. For long wave lengths, *i.e.*, for a frequency range such that $h\nu << kT$, we may expand the exponential in Eq. (36) and neglect powers of $h\nu/kT$ higher than the first. We have

$$e^{\frac{h\nu}{kT}} = 1 + \frac{h\nu}{kT} + \frac{(h\nu)^2}{2!(kT)^2} +$$

so that very nearly

$$e^{\frac{h\nu}{kT}} - 1 = \frac{h\nu}{kT}$$

Using this value in Eq. (36), there follows

$$u_\nu = \frac{8\pi\nu^2}{c^3} \cdot kT \tag{38}$$

which is valid for the long wave-length portion of the spectrum. It is of interest to note that this is the law predicted by classical theory for the whole spectral range and is known as the *Rayleigh-Jeans law*. It is obviously an impossible law, since it predicts an infinite energy density at any finite temperature.

On the other hand, for the high frequency, short wave-length region of the spectrum in which $h\nu >> kT$, we may set

$$e^{\frac{h\nu}{kT}} - 1 \cong e^{\frac{h\nu}{kT}}$$

and obtain

$$u_\nu = \frac{8\pi h\nu^3}{c^3} e^{-\frac{h\nu}{kT}} \tag{39}$$

This is the form of the radiation law obtained by *Wien* by semi-empirical methods and is known by his name. It is more convenient than the Planck law for calculation purposes in the short wave-length region of the spectrum.

The Stefan-Boltzmann law may be derived from Planck's law by performing the integration indicated by Eq. (34). We have

$$u = \frac{8\pi h}{c^3} \int_0^\infty \frac{\nu^3\, d\nu}{e^{\frac{h\nu}{kT}} - 1}$$

and, if we set $h\nu/kT = x$, we find

$$d\nu = \frac{kT}{h} dx$$

$$\nu^3 = \left(\frac{kT}{h}\right)^3 x^3$$

and substituting

$$u = \frac{8\pi k^4 T^4}{c^3 h^3} \int_0^\infty \frac{x^3\, dx}{e^x - 1} = \alpha T^4 \tag{40}$$

where

$$\alpha = \frac{8\pi k^4}{c^3 h^3} \int_0^\infty \frac{x^3\,dx}{e^x - 1}$$

The integral has the value $\pi^4/15$, so that

$$\alpha = \frac{8\pi^5 k^4}{15 c^3 h^3}$$

Finally, the position of the maximum of any one of the curves of Fig. 238 may be obtained by differentiating Eq. (37). The result may be expressed in the form

$$\lambda_m T = \text{constant} = b \tag{41}$$

where λ_m is the wave length for which E_λ is a maximum and the constant b has the numerical value

$$b = 0.288(\text{cm.-°C.})$$

This important law is known as *Wien's displacement law* and was derived by Wien with the help of thermodynamic considerations.

93. Photometric Units; Visibility of Radiant Energy.—Photometry is the study of the measurement of radiant energy in the visible region of the spectrum and, although bearing an intimate relation to the general theory of radiation discussed in the preceding sections, it possesses its own peculiar units which differ somewhat from those which we have been employing. There are two reasons for the unique procedure adopted when dealing with visible radiation: (1) Photometric units were introduced independently of those employed in other branches of physics, very much as the calorie is introduced in heat as an energy unit; and (2) the human eye is not equally sensitive to radiations of different wave lengths (even in the visible range), so that the so-called *relative visibility* of radiation must be included in the definition of photometric units.

Suppose we consider an experiment in which an observer looks at a number of essentially monochromatic sources. The energy radiated from the sources is varied so that they all seem equally bright to a normal observer. It is found that the least radiant power (E_λ) is needed for the source of wave length 5,550 A. The ratio of E_λ for this source to its value for any other source at wave length λ to produce the sensation

of equal brightness for a normal eye is called the *relative visibility* of this wave length. Figure 239 is a plot of this relative visibility $v(\lambda)$ as a function of λ for a normal eye. Because of this selective action of the eye, a statement of the emissive power E in absolute units of a light source is insufficient to determine its visual effect. For this reason it is conventional to introduce a photometric unit of the total rate of emission of energy (energy flux) called the *lumen*. For a normal observer it is equivalent to $\frac{1}{621}$ watt at the wave length 5,550 A. The numerical

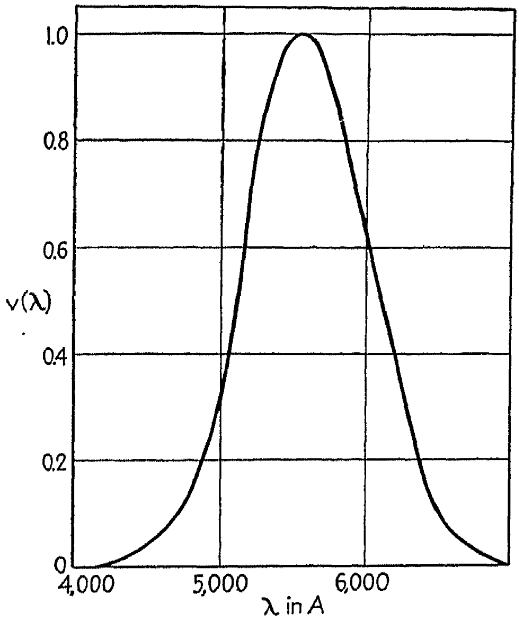

FIG. 239.

factor of 621 arises from the arbitrary definition of a unit of light intensity (see below), the *candle*. Thus 1 watt radiated at 5,550 A. equals 621 lumens of light flux, and at any other wave length the luminous flux F_l in lumens is given by

$$F_l = v(\lambda) \cdot 621F \tag{42}$$

where F is the radiated power in watts from a monochromatic source of wave length λ, and $v(\lambda)$ is the relative visibility of this wave length. The ratio F_l/F is called the *luminous efficiency* of the source. If the source is not monochromatic, then one must integrate over the spectrum to obtain the luminous efficiency. The number of lumens radiated per unit area of the surface of the source is called the *luminosity* of the source, L.

It is clear that the contribution to L from the wave-length range $d\lambda$ is given by

$$dL = 621v(\lambda)E_\lambda \, d\lambda$$

where $E_\lambda \, d\lambda$ is in watts per square centimeter, so that the luminosity becomes

$$L = 621\int_0^\infty v(\lambda)E_\lambda \, d\lambda \tag{43}$$

in lumens per square centimeter, and the luminous efficiency is given by

$$\text{Eff.} = 621\frac{\int_0^\infty v(\lambda)E_\lambda \, d\lambda}{\int_0^\infty E_\lambda \, d\lambda} \tag{44}$$

since $\int_0^\infty E_\lambda \, d\lambda$ is equal to E, the emissive power of the surface in watts per square centimeter.

The *intensity* I of a light source of dimensions sufficiently small that it may be considered a point source is defined as follows: Consider a pencil of rays coming from the source of solid angular opening $d\omega$ (Fig. 240). Then the intensity of the source in the direction n is defined as the luminous flux crossing the area dA' divided by the solid angle $d\omega$. Thus we have

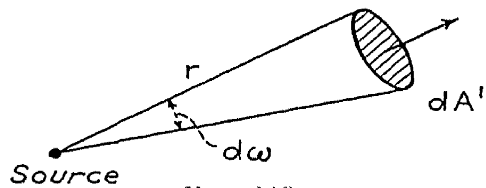

FIG. 240.

$$I = \frac{dF_l}{d\omega} \tag{45}$$

and, if the source radiates uniformly in all directions,

$$I = \frac{F_l}{4\pi} \tag{46}$$

where F_l is the total flux coming from the source. The photometric unit of intensity is an arbitrary unit, the *candle*, defined as a source which (emitting uniformly in all directions) emits 4π lumens, so that 1 candle equals 1 lumen per unit solid angle. Actually the standard candle was chosen arbitrarily, the unit flux obtained from Eq. (46). Experiment then yielded the numerical factor 621 used above.

Consider the area dA' of Fig. 240. The illumination E' on this surface is defined as the ratio of the luminous flux incident on it to its area. Thus

$$E' = \frac{dF_l}{dA'} = \frac{dF_l}{d\omega} \cdot \frac{d\omega}{dA'} = \frac{I}{r^2} \tag{47}$$

using Eq. (45) and the fact that $dA' = r^2\, d\omega$ for the case in which dA' is normal to r. If the normal to the surface element makes an angle θ with the direction of the pencil, then Eq. (47) evidently becomes

$$E' = \frac{I}{r^2} \cos \tag{48}$$

Finally, the brightness K of a pencil of rays has already been defined by Eq. (1). The same definition is utilized in photometry except for a change in units. If, in Eq. (1), one expresses dF in lumens, then we have

$$K = \frac{\Delta I}{\cos\theta\, \Delta A} \tag{49}$$

where $\Delta I = dF/d\omega$ in candles, and K is measured in candles per square centimeter. Equation (49) gives the brightness of the radiation at the surface ΔA. If the surface element ΔA is self-luminous, then K is called the *brightness of the surface.* If such a surface obeys Lambert's law [Eq. (12) or (13)], it is convenient to define a new unit of brightness B, the *lambert,* defined by the relation

$$B = \pi K \tag{50}$$

where K is in candles per square centimeter. Thus the lambert is $1/\pi$ times as large as the candle per square centimeter.

In terms of this new unit, Eq. (13) becomes

$$\bar{E} = E' = B \tag{51}$$

since, for a black surface, the illumination (energy incident per unit time per unit area) equals the emissive power $\bar{E}$. For surfaces which diffusely reflect or transmit light and obey a law similar to Eq. (10) for the reflected or transmitted light (this is essentially Lambert's law), one writes for the surface brightness

$$B = kE' \tag{52}$$

where k is the fraction of the incident light reflected or transmitted and E' is the illumination of the surface.

Problems

1. Consider a black body of surface area A, the surface of which is everywhere convex, so that none of the radiation leaving any point of this surface impinges on any other point of the surface. If this body is completely surrounded by an enclosure and both the body and the walls of the enclosure are at absolute temperature T_1, show that the rate at which radiant energy falls on the body is given by

$$\sigma A T_1^4$$

independent of the area and nature of the walls.

Using the above result, show that, if the temperature of the black body is maintained at T_2 and that of the surrounding walls at T_1, the net rate of gain or loss of energy by radiation of the black body is given by

$$A\sigma(T_2^4 - T_1^4)$$

What condition must be satisfied in order that this law be valid?

2. Taking an average temperature of the entire earth's surface as 15°C. and assuming the earth as a whole (including the atmosphere) to absorb the sun's radiation like a black body and to radiate like a black body, compute the absolute temperature of space (*i.e.*, of the radiation in empty space around the earth). The solar energy falling on the earth's surface is to be taken as 2 cal./cm.^{2}-min., when the sun is directly overhead.

3. A closed graphite crucible at 27°C. is placed inside a furnace whose walls are maintained at a temperature of 1730°C. Treating the crucible as a black body of surface area 40 cm.^{2}, compute the initial rate at which the crucible gains heat by radiation from the furnace walls, using the result of Prob. 1.

If the crucible has a mass of 100 grams and a mean specific heat of 0.3, how long will it take for the crucible to reach a temperature of 100°C.? (Note that the ratio of crucible to wall temperature is small compared to unity for the temperature range in question.)

4. A small hole of area ΔA is made in the walls of a furnace containing black radiation at temperature T. A black sphere of radius r is placed in front of the hole at a distance R from it. Neglecting the radiation from the furnace walls, show that the energy falling on the black sphere per unit time is given by

$$\sigma \Delta A\, T^4 \cdot \frac{r^2}{R^2}$$

5. A black sphere of 8 cm. diameter is placed in front of a small opening (area = 10 cm.^{2}) in the walls of a furnace with its center 40 cm. from the opening. The walls of the furnace are so shielded that only the radiation from the opening is incident on the sphere. The steady temperature reached by the sphere is the same as that which can be maintained by supplying 2.80 watts of electrical power to a heating unit inside the sphere, when the hole in the furnace walls is covered.

Compute the temperature of the furnace radiation, assuming it to be black.

6. Prove that the net rate of heat transfer per unit area by radiation between two plane parallel surfaces of separation small compared with the linear dimensions of the surfaces is given by

$$\frac{\sigma(T_1^4 - T_2^4)}{\frac{1}{\epsilon_1} + \frac{1}{\epsilon_2} - 1}$$

where ϵ_1, T_1, and ϵ_2, T_2 are the emissivities and absolute temperatures of the two surfaces. σ is the Stefan-Boltzmann constant.

7. The temperature of the water in a thermos bottle is observed to fall from 100 to 99°C. in 30 min. when the outer shell of the bottle is at 25°C. Find the time required for the temperature of the same amount of ice water at 0°C. in the bottle to rise to 1°C. if the outer shell is at 20°C. Neglect the heat loss through the cork stopper.

8. The inner bottle of a thermos bottle (unsilvered) contains 1 liter of ice water and the outer shell is at 25°C. The emissivity of a glass surface is 0.85, and the area of the outer surface of the inner bottle (and also the inner surface of the outer shell) is 175 cm². Neglecting heat losses through the cork, find how long it takes for the water's temperature to rise from 0°C:

a. To 1°C.

b. To 10°C. (Use reasonable approximation methods to obtain the answer to part *b*.)

9. Prove that Wien's displacement law follows from the Planck formula (36) or (37) or from Eq. (39).

10. The wave length of maximum intensity in the solar spectrum is 5,000 A. Assuming the sun to radiate like a black body, compute its surface temperature.

11. To what temperature would the blackened spherical bulb of a thermometer rise in full sunlight if the bulb were surrounded by a perfectly transparent, evacuated glass bulb? The surroundings are at 25°C. Assume the thermometer bulb to behave like a black body.

12. Compute the ratio of the increase of brightness of black-body radiation at a wave length of 6,410 A. for an increase of temperature from 1200 to 1500°abs.

13. Starting from Planck's law or from Wien's law [Eq. (39)], prove that the maximum value of the emissive power *per unit wave-length range* (E_λ) for black-body radiation varies with the fifth power of the absolute temperature.

What is the corresponding law for E_ν?

14. What is the radiation pressure of black-body radiation at a temperature of 6000°abs.?

15. Starting from Planck's law, compute the value of b in Wien's displacement law [Eq. (41)].

Repeat the calculation starting from Wien's law [Eq. (39)], and compare your result with the experimental value.

16. Find the brightness of a sheet of paper which is placed on a desk at a distance of 1 meter below a 250 cp. point source. The paper diffusely reflects 80 per cent of the light incident upon it.

INDEX

F